A VALIDATION STUDY OF SEVERAL GERIATRIC BEHAVIOR RATING SCALES

By

PAUL K. CHAFETZ

A DISSERTATION PRESENTED TO THE GRADUATE COUNCIL
OF THE UNIVERSITY OF FLORIDA IN
PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1981

COPYRIGHT 1981

BY

PAUL K. CHAFETZ

This dissertation is dedicated to my parents, Irving and Mildred Chafetz, whose love for me and support of my chosen pursuits have been unfailing.

This dissertation also represents the completion of a project they undertook twenty years ago--the higher education of their four children. Their dreams and sacrifices for us are lovingly acknowledged.

ACKNOWLEGEMENTS

This project, which spanned two years, was made possible by the guidance, assistance, and cooperation of several people. Foremost among these are my committee members. By way of this project and others, my chairman, Walter Cunningham, devoted hundreds of hours to my development both as a scientist in particular and a professional in general. His talents, time, and concern have been very much appreciated.

Warren Rice, as my only committee member located at the research site, provided priceless assistance with the many logistical difficulties involved in any research project, as well as providing significant scholarly direction to the project. Without his help, especially when I was away on internship, completing this project would have been inestimably more difficult.

Drs. Gordon Streib, Suzanne Johnson, and Cynthia Belar, by sharing their expertise with me, helped me greatly in my efforts to design and implement a project that was clinically relevant and scientifically interesting.

The many professionals and family members who served as raters are to be robustly thanked for their energies. The nursing staff on the neurology ward of the Gainesville Verterans Administration Medical Center, under the leadership of Head Nurse Betty Polotsy, deserve special mention. The psychology staff of that hospital, who tolerated my frequent use of valuable office space, are also appreciated.

The University of Florida Center for Gerontological Studies contributed materially to my progress on this project in the form of a research assistantship and many semesters of tuition support.

Finally, the Division of Medical Psychology at Duke University Medical Center, and the staff of the computational laboratory of the Duke University Center for the Study of Aging and Human Development, Durham, North Carolina, graciously provided computational support and facilities for this project.

TABLE OF CONTENTS

CHAPTER		PAGE
	ACKNOWLEDGEMENTS	iv
	LIST OF TABLES	vii
	ABSTRACT	viii
I	INTRODUCTION	1
	Objective	1
	Background	1
	Psychometric Criteria for Outcome Measures	3
	Geriatric Treatment Program Outcome Research	6
	Reality Orientation Therapy Outcome Research	8
	Purpose of the Study	13
II	METHOD	14
	Measures	14
	Subjects	19
	Procedure	21
	Interrater Reliability	21
	Construct Validity	22
	Criterion-Related Validity	24
	Expectations of Scale Performance	26
III	RESULTS	29
	Interrater Reliability	29
	Construct Validity	29
	Criterion-Related Validity	32
	Relative Discriminative Power of the Scales	42
IV	DISCUSSION	46
	Review of Findings	46
	Limitations and Biases of the Study.	49
	Choosing a Scale	54
	Subject-Absent vs. Subject-Present Scales	55
	Summary of Findings on Each Scale	56
	Further Research	58
	Significance of the Study	61
APPENDICES		
A	SUBJECT INFORMED CONSENT FORM	62
B	SCALES	64
C	PRACTICAL PROBLEMS OF DOING RESEARCH	77
REFERENCES		81
BIOGRAPHICAL SKETCH		86

LIST OF TABLES

TABLE		PAGE
ONE	INTERRATER RELIABILITY	30
TWO	GROUP MEANS AND STANDARD DEVIATIONS	31
THREE	CONVERGENT CONSTRUCT VALIDITY: CORRELATION MATRIX	33
FOUR	CONVERGENT CONSTRUCT VALIDITY: PRINCIPAL COMPONENTS ANALYSIS	34
FIVE	CRITERION-RELATED VALIDITY: COMPARISON ONE	35
SIX	CRITERION-RELATED VALIDITY: COMPARISON TWO	37
SEVEN	A POSTERIORI MULTIPLE COMPARISONS OF GROUPS I, III, AND IV	39
EIGHT	CRITERION-RELATED VALIDITY: COMPARISON THREE	40
NINE	A POSTERIORI MULTIPLE COMPARISON OF ALL SUBJECT GROUPS	41
TEN	CRITERION-RELATED VALIDITY: COMPARISON FOUR	43
ELEVEN	CLASSIFICATION HIT RATES, GROUPS I AND IV	45

Abstract of Dissertation Presented to the Graduate Council of the University of Florida in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

A VALIDATION STUDY OF SEVERAL GERIATRIC BEHAVIOR RATING SCALES

By

Paul K. Chafetz

March, 1981

Chairman: Walter R. Cunningham, Ph.D.
Major Department: Department of Clinical Psychology

Five scales designed to measure Activity of Daily Living skills in geriatric patients were studied: the Physical Self-Maintenance Scale, the Rapid Disability Rating Scale, Factor 3 of the Geriatric Rating Scale, The Minimal Social Behavior Scale, and the Performance Test of Activities of Daily Living. The first three are subject-absent scales, meaning that raters complete the scales based on their knowledge of the subject. The last two are subject-present scales, in that they are directly administered to each subject individually. The interrater reliability (IRR), construct validity, and concurrent criterion-related validity of each was evaluated. Hospitalized male veterans (N=110) on neurology, medicine, and psychiatry wards, and nonhospitalized male veterans (N=30) were subjects.

Dual ratings of 50 subjects yielded excellent correlational IRR results, ranging from .853 to .985, all significant at p=.0001. Construct validity of the scales proved very strong through both correlational and factor analytic techniques. The two-scale correlations ranged from .498 to .893 in absolute value and were all significant at p=.0001. A principal components analysis yielded one factor which accounted for 79% of the total variance, and the scale loadings in this factor ranged from .942 to .793.

Each scale's criterion-related validity along each of the following dimensions was evaluated: (1) age among male neurology inpatients, (2) type of impairment among elderly male inpatients, within broad categories, (3) presence of neurological pathology among males, and (4) inpatient status among elderly males. For each criterion, the appropriate subsamples were grouped and a Hotelling's t-squared analysis and least significant difference post-hoc multiple comparison conducted. With just one scale-by-criterion exception, every scale clearly demonstrated validity along all four dimensions. The hit rate of scales in discriminating elderly neurological and psychiatric patients ranged from 65.00% to 80.85%.

Several methodological difficulties emerged during this study. First, subject recruitment procedures were less systematic and rigorous in some samples than others and may have caused slight bias in some sample means. Second, many subjects had two or three of the subject-absent scales completed by the same rater, bringing into question the adequate independence of these ratings. With non-hospitalized subjects, more rigorous rater training would have been desirable. Third, the participation rate of non-whites in the study varied widely between samples, from 23.3% to 0.0%. However, no scale correlated significantly with race for the entire sample, and only one scale by race correlation was significant among the individual samples.

The highly favorable psychometric results with every scale evaluated leave potential test users with choices of both method (i.e., subject-absent vs. subject-present) and specific instruments. Subject-absent scales are probably the technology of choice (1) whenever many

indigenous raters are available and trainable, (2) when there may be a lack of subject motivation to be assessed, and (3) especially when retesting of subjects is planned or desirable. This will often be the case in residential institutions. Subject-present scales, on the other hand, seem most appropriate for research and screening purposes involving one-time evaluation of subjects. The present data recommend the Physical Self-Maintenance Scale and the Performance Test of Activities of Daily Living most highly.

Future research should emphasize the collection of good normative data with these scales and on extending studies like the present one to female populations.

CHAPTER I
INTRODUCTION

Objective

The overall objective of the proposed study was to promote the quality of future gerontological research by identifying some behavioral measurement instruments which possess adequate psychometric characteristics to justify continued use. Only scales which are clearly applicable to use with neurologically impaired geriatric patients were chosen. The project clarifies the appropriateness of the scales for both individual assessment of such patients and treatment outcome research with groups of such patients.

Background

During the twentieth century, there has been considerable progress on such social frontiers as medical science, medical technology, and occupational safety. This has been accompanied by a dramatic growth in the ranks of the elderly, both in terms of their numbers and as a percentage of the American population. Between 1870 and 1970, the number of Americans age 65 and over increased from 1.2 million to approximately 20 million. The portion of the total population they comprise grew from 2.9% to 9.9%. Estimates of the number of elderly Americans in the year 2000 range from 28.8 million (Brotman, 1973) to 33 million (Butler, 1969). The proportion of elderly veterans is also growing. In 1975, 13% of American veterans were age 60 or over, but this statistic has been projected to reach 39% by the year 2000 (Kirchoff, 1977).

While the majority of elderly persons are in relatively good health and are functioning independently, advancing age is still associated with decline of function. Of those 75 and over, 23.7% are unable to carry on major activities, as opposed to 9.7% of those between 65 and 75 years of age (Busse, 1969). Almost 1% of all those over 65 are in some type of mental institution. An additional 3% are in other types of institutions, such as nursing homes, chronic disease hospitals, and homes for the aged (Redick, Kramer, & Taube, 1973).

Among the roughly one million elderly persons who are institutionalized, organic brain disorders are extremely widespread, though their prevalence in different institutions varies according to certain facility-specific selection factors. Thus, a study of state hospitals in North Carolina indicated that, of those admitted for the first time after age 65, 79% had a diagnosis of acute or chronic brain syndrome (Whanger, 1971), whereas only 42% of elderly residents in private psychiatric facilities had organic brain syndromes (Redick, Kramer, & Taube, 1973).

In response to these population trends and institutional parameters, geriatricians are becoming increasingly innovative in their search for new and effective treatments. Examples range from drugs to planned psychosocial environments. Notable among recent developments are reality orientation therapy (Folsom, 1968), Gerovital (Jarvik & Milne, 1975), and hyperbaric oxygenation (Eisner, 1975; Thompson, 1975). The effectiveness, and in some cases even the use, of many recently developed therapies is still quite controversial. These debates, which are carried on by concerned theorists,

administrators, and clinicians by way of scientific journals, should be based on, and ultimately resolved with, sound empirical data.

Psychometric Criteria for Outcome Measures

"Sound" data are those which have been obtained from measurement techniques or instruments of demonstrated reliability and validity. Why is this an important requirement? Use of data-gathering methods without these characteristics creates a situation in which it cannot be said with certainty what the data represent--the intended variable, some systematic intervening variable, experimenter or rater bias, methodological or statistical artifact, or random error. When the results of studies which utilized instruments of unknown reliability or validity are used to formulate policy and plan future investments of institutional resources, the risk of being misled by speciously "empirical" data can be very great. This pitfall is best avoided by choosing dependent measures whose psychometric characteristics are already known to be acceptable. In geriatrics, just as in other fields, the speedy development of effective and efficient health care delivery systems requires a dependable technology for accurate program evaluation.

Reliability refers, in general, to the consistency of scores obtained with a measurement instrument. A test should be able to yield, with a single application, a trustworthy estimate of the "true" level of the studied variable at the time of measurement. The presence of excessive score variability (inconsistency) brings into serious question the test's ability to do this. A sine qua non for scales intended for use in research of any kind, adequate reliability is absolutely essential if the scale is to be sensitive to changes in the measured variable.

A scale's reliability must be estimated by comparing the results of multiple measurements of a sample for agreement. Several methods exist for making such comparisons, each addressing different sources of error variance. Test-retest reliability measures error due to time, but confounds time with possible practice effects, and with real change over time. Alternate forms of many tests are available to help avoid practice effects in test-retest comparisons, but use of alternate forms may present the difficulty of unequal content sampling. By computing split-half reliability on data gathered within a test situation, the effect of content sampling can be isolated. Interrater reliability is the agreement of two or more observers who simultaneously score the same sequence of events. High interrater reliability suggests that the scores obtained by one rater would be very similar to scores obtained by any other comparably trained rater(s) who had rated the same events. In such cases, duplication of effort (through multiple rating) is less necessary, thus making assessment more efficient and accurate. Interrater reliability has been termed "the primary reliability concern in applied behavior research" (Walls, Werner, Bacon, & Zane, 1977, p. 83).

Validity, like reliability, is not a unitary concept. It refers, in general, to the extent to which a test actually measures the particular construct it is supposed to measure. It is also related to the amount of confidence with which a test user may apply an individual's test score in making a decision about that individual. It is important to remember, however, that questions of a scale's validity can be meaningfully explored only after adequate reliability has been demonstrated. After reliability has been achieved, then

begins the slower process of assessing the scale's several validities (Standards for Educational and Psychological Tests, 1974).

To demonstrate content validity, one must show that the behaviors elicited in testing constitute a representative sample of all possible behaviors in the specified performance domain. This should not be confused with relatively worthless judgements of appearances of relevance that have in the past been termed "face validity." Construct validity is of two types--convergent and discriminant. Convergent validity is demonstrated when the test scores on a sample of subjects correlate highly with other measures of the same construct made on the same subjects. Discriminant validity requires that the results of this scale vary independently of the subjects' scores on tests designed to measure different constructs or variables (Campbell & Fiske, 1959). There are also two types of criterion-related validity. Concurrent criterion-related validity is the extent to which an individual's test score is useful in estimating his present standing on another variable. Predictive criterion-related validity, however, involves a time element, and indicates the extent to which the individual's future level on some criterion can be predicted from a knowledge of prior test scores. Important in this regard is how much the knowledge of test scores improves prediction to the criterion over simply using population base rates. The magnitude of this improvement, often called incremental validity, depends not only on the power of the test, but also on the behavior's base rate (Meehl & Rosen, 1955). The closer the base rate of a certain characteristic or behavior in a given population is to .5, the easier it will be to construct a test which will significantly improve the clinician's

ability to predict the emergence of that characteristic or behavior. However, prediction of events which are already known to have very low base rates (such as changes of eye color), or very high base rates (such as dying by age 90), will be improved very little by adding test information to base rate information.

Geriatric Treatment Program Outcome Research

A major compilation of rating scales for geriatric research, with an emphasis on psychopharmacological research (Salzman, Kochansky, & Shader, 1972), listed 117 such instruments, including 42 behavior rating scales. Since the publication of that review, dozens of additional scales have appeared, apparently resulting from the burgeoning interest in aging. Unfortunately, data on the validity, and especially on the reliability, of the vast majority of these scales are often either indequate or nonexistent.

It is, of course, the test user's responsibility to marshal evidence in support of any instrument to be used to help make decisions (Standards for Educational and Psychological Tests, 1974, p. 32). Yet, how can the individual gerontological researcher hope to choose wisely from the plethora of available scales? The most recent Mental Measurements Yearbook (Buros, 1972) reviews none of Salzman's (et al., 1972) 42 behavior rating scales, nor, of course, any of those which have appeared since 1972. The Reality Orientation Program at the Tuscaloosa Veterans Administration (VA) Hospital supplied a brief annotated list (List of Research and Evaluation Tools, 1976) of rating scales, but without critical comments or psychometric data. Walls, Werner, Bacon, and Zane (1977) list 166 behavior checklists along with "selective descriptive notes." Many of these scales may be applicable

to elderly populations, although they are not marked as such. Each scale is rated for objectivity, and the availability of reliability and validity information is described (i.e., yes or no). Actual information as to the reliability and validity of each scale, however, is not included.

Many clinicians and researchers have need of good dependent measures for geriatric program research and clinical assessment. Most, however, probably do not have the time, resources, or training to search out and evaluate what reliability and validity relationships have been demonstrated. Also, it would be quite wasteful for each interested party to be forced to perform the same search and evaluation as others before them. The inevitable product of this unfortunate situation is that much serious research is conducted using measures of unknown dependability. Many choices of measurement instruments are certainly haphazard, and this can only lead to haphazard findings.

It may be argued that accurate measurement of discrete, well-defined behaviors is all the user should require of a behavioral rating scale, and that the exploration of sophisticated questions of reliability and validity is either superfluous or inappropriate. After all, the correlation of Scale X with Scale Y may be of little interest as long as Scale X data can indicate a patient's ability to, for example, wash his face. Indeed, if the user is truly interested only in the precise behaviors actually rated, validation efforts are quite unnecessary (although reliability is still required). While such usage of behavioral rating scales can be perfectly legitimate, however, it is certain that behavioral rating scales of the type

discussed here are usually used in order to obtain an estimate of a patient's competence in certain areas of performance. Each "area" subsumes many discrete, but functionally related, motor behaviors. While any reliable scale can supply accurate information about the specifically rated behaviors, this information can be used to make inferences about the patient's competence in that class of behaviors only if the scale's validity has been demonstrated.

Reality Orientation Therapy Outcome Research

One body of literature in which use of inadequate dependent measures is all too common is reality orientation therapy (ROT) outcome research. Five such studies will be reviewed here to illustrate the problems discussed above. ROT was developed at the Topeka and Tuscaloosa Veterans Administration (VA) hospitals for the treatment of geriatric inpatients with a moderate to severe degree of memory loss, confusion, and disorientation. It is designed to combat the process of cognitive deterioration in two ways--the patient is continually stimulated through the repeated presentation of fundamental information, and he is placed in a group where he meets and competes with other patients, thus forcing him out of his isolation and back into his environment (Stephens, 1970).

Letcher, Peterson, and Scarbrough (1974) devised a four level classification of nursing care requirement to assess the effectiveness of the geriatric rehabilitation program at the Tuscaloosa VA hospital. Although 24-hour and classroom ROT was a central element of this program, the absence of control data made it impossible to isolate the role of ROT from that of other program components. Subjects were 125 male inpatients, whose mean age was 82.8 years. Eighty percent were

diagnosed as having organic brain syndrome secondary to cerebral arteriosclerosis, and another 15% had suffered strokes. All subjects also had at least one additional major medical problem. Treatment duration, which was recorded but not manipulated or controlled, ranged from 2 to 56 months. The "experiment" involved no control groups, no assessment of the classification scheme's reliability, and no tests of significance. Rather, percentages of the sample which required each level of nursing care before and after the individual's ROT term are supplied. That is, pre-ROT evaluation by the treatment team rated .8% of the sample at level I (minimal care), 8.0% at level II, 23.2% at level III, and 68.0% at level IV (total care). Corresponding post-ROT figures were 12.0% 14.4%, 24.8%, and 48.8%. Although the authors confidently concluded that "on the basis of the current data, reality orientation as part of a total rehabilitation program can be evaluated as highly successful" (p. 803), it is clear that, in the absence of reliability information, control data, or appropriate tests of statistical significance, such claims were premature and poorly grounded.

Barnes (1974) studied two male and four female nursing home residents with a mean age of 81 years. All six were diagnosed as having senile dementia, exhibiting confusion, memory loss, and disorientation. Subjects were assessed four times--initially, after a four-month control period, immediately following an intensive, six-week ROT treatment period, and one week after termination of ROT. Data were obtained with a questionnaire consisting of (a) 23 questions for each subject to answer as a measure of learning, and (b) five questions about the subject's behavior, to be completed by the nursing

director. Despite the author's acknowledgement that the reliability and validity of the questionnaire were not assessed, he was not deterred from administering it repeatedly, interpreting its results, and submitting them for publication. Results of the questionnaire indicated insignificant changes during the control and treatment periods in the subjects' learning and ward behavior. Follow-up testing after one week showed a significant decline in the subjects' learning but no behavioral changes. In support of ROT effectiveness, however, the author then resorted to nurses' anecdotal reports, which were highly favorable. In fact, it was quite meaningless to infer anything about the effectiveness of ROT from the results of a measure with unknown reliability and validity. It appears that, at best, the questionnaire may have been insensitive to the positive changes which the nurses claimed to have observed.

Brook, Degun, and Mather (1975) studied eight male and ten female psychogeriatric inpatients. The patients, whose mean age was 73.3 years, had been continuously hospitalized for an average of 1.9 years. Subjects were assessed using an undescribed "rating scale of intellectual and social functioning." The raters and rating procedure were not described, and factors such as reliability and validity were not even mentioned. On the basis of initial ratings, subjects were divided into three groups according to level of functioning. Half of each group served as control subjects. Experimental subjects took part in intensive classroom ROT for eight weeks, while control subjects were simply brought to the ROT room daily and left to occupy themselves. Weekly testing suggested that all subjects made initial gains, presumably due to the novelty (Hawthorne effect). Thereafter,

however, control subjects deteriorated back to baseline, while experimental subjects continued to progress. By the eighth week, within each condition, the groups who were initially rated as "high functioning" were performing slightly better than "middle functioning" groups, who, in turn, bettered the "low functioning" groups. Again, there is no reason to have confidence in these findings and conclusions, since the quality of the data is totally unknown.

Harris and Ivory (1976) compared two groups of 29 female geriatric patients at the Florida State Hospital. The ROT group had a mean age of 66.6 years and had been continuously hospitalized for an average of 24.6 years. The control group averaged 71.1 years of age and 23.0 years of hospitalization. Most subjects carried diagnoses of chronic organic brain syndrome, while a few were mentally deficient. All subjects received traditional hospital treatment, including chemotherapy, music therapy and occupational therapy. Experimental subjects additionally received 24-hour ROT, administered by psychiatric aides. After an unstated period of treatment, all subjects were rated using the Florida State Hospital Patient Behavior Rating Sheet, which addresses a variety of ward behaviors and "verbal orientation" behaviors, as well as recording the rater's own impression of the subject in several behavioral spheres. Reliability data are not supplied. It is curious that, although all subjects were rated simultaneously and independently by two raters, and eight pairs of raters participated in the study, interrater reliability was not reported. Also disconcerting is the fact that each aide who worked with a subject was also a rater, and all aides, or course, knew which patients were assigned to each group. This already enormous risk of

rater bias (O'Leary, Kent, & Kanowitz, 1975) was amplified by part of the rating procedure, which required the aide/rater to elicit and rate scorable responses. Parenthetically, no treatment effect on ward behavior emerged, although ROT subject performed significantly better on six of nine verbal orientation tasks.

Citrin and Dixon (1977) reported perhaps the most rigorous attempt to assess the effectiveness of ROT. At a relatively large geriatric institution in Lincoln, Nebraska, ROT was instituted on one ward, and withheld from another. The comparability of the two wards was not described, however, and it is, therefore, impossible to know whether assignment of patients to wards (and, therefore, to treatments) was random. Twelve patients of the ROT ward, mean age 84 years, and 13 patients on the control ward, mean age 83 years, served as subjects. All showed disorientation, but none were nonambulatory, blind, or deaf. Two measures were administered to all subjects before and after a two-month period, during which the experimental subjects received 24-hour and classroom ROT. One measure used was the Geriatric Rating Scle (GRS) (Plutchik et al., 1970), a well established, 31-item behavioral scale whose reported reliability ranges from .87 to .94 (see CHAPTER II below). Also used was the unpublished Reality Orientation Information Sheet (ROIS), developed in 1973 at the Tuscaloosa V.A. hospital. The ROIS includes 20 orientation questions to be asked of the subject and five opinion questions to be asked of the ward aide about the subject. These authors obtained an interrater reliability of .98 with the ROIS, based on a sample of ten patients. Experimental data were reduced to four scores for each test--mean pre- and post-treatment scores for the experi-

mental and control groups. An elaborate, and wholly inappropriate (Campbell & Stanley, 1963, pp. 22-23) array of t-tests was unfortunately used to test for treatment effects, rather than computing gain scores (post-treatment scores minus pre-treatment scores), upon which t-tests could be meaningfully conducted. The array method yielded "significant" results with the ROIS data but not with the GRS data, from which it was concluded that ROT may have a positive impact upon cognitive orientation but little effect upon behavioral functioning.

Purpose of the Study

In response to the unfortunate quesswork now involved in choosing a behavioral dependent measure for geriatric therapy outcome research, the present study directly compared a small number of those scales which showed the most promise for successful refinement and popular acceptance. Reliability and validity were rigorously assessed, and the advantages and disadvantages peculiar to each scale were identified. On the basis of this newly generated information, straightforward recommendations of measures for specific purposes are made.

CHAPTER II
METHOD

Measures

As mentioned above, dozens of behavioral rating scales have been published, many of which are potentially applicable to geriatric research. In order to select a practical number of scales to be studied from the large number available, broad guidelines were developed which proved helpful, but neither exhaustive nor absolute. The major selection criterion specified scales designed to assess an individual's level of function in behavioral domains necessary for normal, independent living. Such domains, often called activities of daily living (ADL), include eating, dressing, grooming, ambulation, use of appliances, etc. The Bender-Gestalt Test (Bender, 1938), by way of contrast, did not meet this criterion. Although it is well known as a diagnostic instrument and is of demonstrated usefulness in differentiating between diagnostic groups (Lacks, Harrow, Colbert, & Levine, 1970), it does not directly address behavioral competence in daily functioning. A more direct, behavioral assessment technology offers the advantage of immediate operationalization of concepts, and clear, clinically useful therapeutic inferences from identified deficits (Lindsley, 1964; Rebok & Hoyer, 1977). It has been demonstrated, as well, that behavioral assessment, when compared with global, subjective ratings, is relatively resistant to raters'

expectancy biases (Kent, O'Leary, Diament, & Dietz, 1974; Schuller & McNamara, 1976).

Scales to be evaluated had to appear capable of measuring highly discrepant levels of functional competence, from normal to "senile" behavior. Procedure length must not be prohibitive, as the Camelot Behavioral Checklist (Foster, 1974), with 399 items, or even the Geriatric Profile (Evenson, 1971), with 74. Available psychometric data had to be at worst, equivocal. Of course, truly damaging data are seldom published. Finally, there had to be no indication that the scale had already met with overt rejection by potential users.

The five scales selected for evalation by these criteria were the Geriatric Rating Scale, the Physical Self-Maintenance Scale, the Rapid Disability Rating Scale, the Minimal Social Behavior Scale, and the Performance Test of Activities of Daily Living.

The Geriatric Rating Scale (GRS) (Plutchik, Conte, Lieberman, Bakur, Grossman, & Lehrman, 1970), a revision of the Stockton GRS (Meer & Baker, 1966), consists of 31 items, each of which is scored 0, 1, or 2. Excellent correlational interrater reliability has been obtained, ranging from .87 (N=86; Plutchik et al., 1970) to .94 (N=20; Plutchik & Conte, 1972), and the correlation between ratings obtained approximately one year apart was .65 (N=132; Plutchik & Conte, 1972). Scores on the GRS differentiated between hospitalized geriatric and hospitalized non-geriatric patients (N=243, $p<.001$; Plutchik et al., 1970) and between organically and functionally impaired psychogeriatric patients (N=73, $p<.001$; Dastoor, Norton, Boillat, Minty, Papadopoulou, & Muller, 1975). Scores have shown high agreement with clinical staff ratings (Plutchik et al., 1970), and may have prog-

nostic value in terms of discharge (Plutchik & Conte, 1972). As a result, the GRS was adopted by the Psychopharmacology Research branch of NIMH for inclusion in its Assessment Battery for Nurses, and one study (Miller & Parachek, 1974) employed the GRS as the criterion measure against which to validate a newer geriatric rating scale.

The GRS has also been factor analyzed (N=370; Smith, Bright, & McClosky, 1977), yielding factors which have been named Withdrawal/ Apathy, Antisocial Disruptive Behavior, and Deficits in Activities of Daily Living. Findings of significant sex differences on the latter two factors only have helped to explain some of the equivocal sex results of earlier research. Significant sex differences also mean that the present project, which studied only males, yielded information generalizable only to males. In the present study, the three GRS factors, as described in Smith et al. (1977), served as autonomous units of analysis in lieu of an overall GRS score.

Based on the advanced psychometric development of the GRS (a quality which the reader will better appreciate after reviewing the following scales), it was expected that this scale would fare quite well in the present study.

The Physical Self-Maintenance Scale (PSMS), adapted by Lawton and Brody (Lawton, 1971, 1972) from a scale used at Langley-Porter (Lowenthal, 1964), is a Guttman-scaled set of observer ratings of competence in dressing, grooming, eating, bathing, locomotion, and toileting. An individual's score ranges from zero to six, one point being given for each function in which the most independent rating is checked. It was considered that an enlarged scoring procedure might make this scale more sensitive to degrees of impairment. Lawton and

Brody (1969) reported that independent ratings of 36 patients by two licensed practical nurses correlated .87, and ratings of 14 other patients by two research assistants correlated .91. Extensive validity information for this scale has not been reported, but data are available differentiating mentally impaired aged from healthier applicants to a home for the aged (Lawton, 1972). Surveying data from 343 cases, Lawton (1971) found mean scores to drop quickly as one moves from old age home applicants (4.1), to elderly mental hospital admissions (3.5), to protective custody patients (2.0).

The Rapid Disability Rating Scale (RDRS) (Linn, 1967) applies a three-point "frequency of occurrence or degree of severity" scale to the assessment of performance in sixteen categories of ADL and ward behavior. Scores thus range from 16 to 48, higher scores reflecting greater disability. Nurse completion of the RDRS for one patient is reported to require only two minutes. Reported interrater reliability estimates range from .894 (N=1000; Linn, Gurel, & Linn, 1977) to .913 (N=20; Linn, 1967). Based on a sample of 1000 male nursing home residents, normative data (mean score and percentage of sample showing the disability) are available for each of the scale's sixteen items (Linn et al., 1977). RDRS scores have been found to correlate significantly with the number of previous hospitalizations, the subsequent death rate, and with physicians' ratings of medical status change in elderly patients. The RDRS has been used frequently by the author and her associates in exploratory research on nursing home program effectiveness and nursing home population characteristics (Carmichael & Linn, 1974; Greenwald & Linn, 1972; Linn & Gurel, 1969; Orgren & Linn, 1971).

The Minimal Social Behavior Scale (MSBS) (Lawton, 1971, 1972) is an adaptation of a scale of the same name by Farina, Arenberg, and Guskin (1957). It requires administration by a clinician and presents the subject with twelve basic social stimuli, such as a greeting and a chance to do a favor, the responses to which are scored for social appropriateness. One of the 25 items which make up this scale involved administering the entire Mental Status Questionnaire (Kahn, Goldfarb, Pollack, & Peck, 1960). The geriatric norms available for this scale indicate that "the average score for relatively intact geriatric mental patients [is] 22, whereas the average score for impaired and uncooperative patients [is] 15" (Lawton, 1971, p. 474). The reliability of this scale is unknown, although its 30-item and 32-item precursors (Mangun & Webb, 1956; Farina et al., 1957) reportedly attained interrater reliabilities of .96 (N=15) and .95 (N=35), respectively. Neiditch and White (1976) included Lawton's MSBS in a test battery for repeated assessment of an inpatient psychogeriatric sample (N=36). Reliability was not assessed. Inclusion of the MSBS improved the clinical team's ability to predict five-month outcome substantially, although the Mental Status Questionnaire accounted for most of this gain. This finding was consistent with that of Markson and Levitz (1973). Dastoor (et al., 1975) who also did not assess reliability, found the MSBS to distinguish between organically and functionally impaired geriatric patients.

The Performance Test of ADL (PADL) (Kuriansky, Gurland, Fleiss, & Cowan, 1976; Kuriansky & Gurland, 1976) is unique in that it requires subjects to actually demonstrate for the test administrator selected activities of daily living. Sixteen tasks are included which may be

roughly subsumed as follows: eating, dressing, grooming and toileting, and use of applicances. Each of the sixteen tasks has been divided into its gross muscular components for scoring. For example, "hair combing" consists of four distinct, scorable elements: (a) takes comb in hand, (b) grasps comb properly, (c) brings comb to hair, and (d) makes combing motion. Developed as part of a cross-national, interdisciplinary study of psychogeriatric assessment technology, the PADL has been reported to have high interrater reliability (.902; Kuriansky & Gurland, 1976), but neither the reliability sample size nor the details of administration and computation are reported. Moderate agreement with verbal ratings by family members (N=45, K=.37), and some prognostic utility have also been reported (Kuriansky et al., 1976).

Subjects

This project involved samples of subjects from five subsets of one population, male veterans. Four samples were gathered within the Gainesville, Florida, Veterans Administration Medical Center (GVAH), and one from among veterans living outside the hospital.

Group I consisted of 30 male inpatients, age 60 and over, on the neurology ward of GVAH. The mean age of this group was 69.37 (SD=8.65), and age ranged from 60 to 87. The modal diagnosis was cardiovascular accident (CVA) (45%), followed by seizures and dementia (13% each).

Group II consisted of 30 male neurology patients under the age of 60, on the neurology ward of GVAH. The mean age of this group was 48.83 (SB=12.15), and age ranged from 19 to 59. The modal diagnosis was stroke (CVA or transient ischemic attack) (30%), followed by dementia (10%).

Group III consisted of 30 male patients, age 60 or over, on the medical wards of GVAH, excluding psychiatry, neurology, and surgery. The mean age of this group was 69.63 (SB=8.92), with a range of 60 to 87. The modal diagnosis was mass, either benign or cancerous (43%).

Group IV consisted of 20 male psychiatric inpatients at GVAH, age 60 or over, whose diagnoses did not include organic impairment. The mean age of this group was 63.25 (SD=3.49), and age ranged from 60 to 73. The modal diagnosis was neurotic depression (35%), followed by affective pyschosis (20%) and chronic schizophrenia (15%).

Group V consisted of 30 relatively healthy, noninstitutionalized male veterans, age 60 and over. Retired men who volunteer at GVAH provided the largest group of this sample (43%), followed by veterans visiting outpatient clinics (36%). The mean age of this group was 72.17 (SD=7.70), with a range of 61 to 86.

For Groups I through IV, patients who met the stated criteria became eligible as participants after seven days of residence on the ward. This gave the raters time to become familiar with the subjects. Those eligible patients who agreed to participate signed a consent form (see APPENDIX A). Subject recruitment ceased when data had been collected on the required number of subjects.

Subject samples were defined by hospital service, rather than by specific diagnosis, because these scales are most often used with populations containing a mixture of diagnoses, all falling within some broad category (e.g., medical, psychogeriatric). The disadvantage, that this method yields less information about what determines a given subject's score, is offset by the advantage, that the samples thus obtained are more representative of the populations with which these scales will be used.

Procedure

Interrater Reliability

Subjects

Thirty Group I subjects and five subjects from each of the four remaining groups served as subjects for this phase of the project, for a total of 50 subjects.

Procedure

Every subject was rated twice with each of the five main scales described above (RDRS, PSMS, GRS 3, MSBS, and PADL). Ward nurses, full-time nursing aides, medical students, medical interns, and medical residents served as raters with the first three instruments (GRS, PSMS, RDRS), and the subjects from Groups I through IV. These raters were briefed by the principal investigator (PI) and/or his assistant concerning the scales and their proper use. The dual ratings with each scale were simulatneous and independent, as actual subject presence was unnecessary for the completion of these scales. Subjects from Group V were rated by their children, spouses, and/or close acquaintances, after being familiarized with the scales.

The principal investigator and his assistant served as the rating pair for the remaining two instruments (MSBS, PADL), as these required direct administration. The assistant, Marty Ludwig, was a senior at the University of Florida majoring in psychology, who continued in his role as assistant after graduating. Each subject was administered each of these two scales only once, but his performance on each scale was scored twice, that is, simulatneously and independently by the PI and the assistant. Both the PI and the assistant conducted actual administrations during the course of data collection, which proceeded from June, 1979, to January, 1980.

Analysis

This study is the first assessment of the interrater reliability of any of the included scales by anyone other than the scales' respective authors. Nine scales were analyzed. These included five main scales (RDRS, RSMS, GRS Factor 3, MSBS, & PADL) and four additional scales (PSMSX, a modification of the PSMS; MSQ, the Mental Status Questionnaire in the MSBS; and GRS Factors 1 and 2). For each scale or factor, a product-moment correlation coefficient was computed to measure the agreement between two independent ratings of subjects. This is the test's interrater reliability. Correlational measurement of interrater reliability was employed because item-agreement methods (Repp, Deitz, Boles, Deitz, & Repp, 1976) are unreasonably conservative when used with instruments that assess several different behaviors (Walls et al., 1977). For purposes of research involving only comparisons of group means, a correlation of .7 is considered good, and findings exceeding .85 are excellent. When a test score is to be used in individual assessment and clinical decision-making, however, higher standards of reliability are called for. Therefore, only those scales which demonstrated interrater reliability of .8 or greater were included in later portions of this study.

Construct Validity

Subjects

The remainder of Groups II through V were then recruited and assessed during the period of January, 1980, to June, 1980.

Procedure

Data were generated with the remaining 90 subjects in essentially the same manner as with the initial 50 subjects, but with one important

change. Each subject was rated by only one rater, rather than two, the scales' interrater reliability having already been established.

Analysis

All data (N=140) collected with scales which reached minimum reliability standards were reanalyzed and a correlation matrix computed to indicate the extent of agreement between scales.* This provides quantitative evidence for the extent of agreement between the scales, or the scales' convergent constuct validity (Campbell & Fiske, 1959).# The correlation values in this matrix were evaluated for statistical significance utilizing an alpha level of .05.

It was to be hoped that scales purporting to measure competence in highly similar behavioral domains would produce highly correlated scores. To the degree that results conform to this expectation, the convergent construct validity of the correlated scales is supported. It was expected, specifically, that the MSBS, which is designed to be sensitive to a slightly higher level of function than the other scales, would correlate less highly with the other scales than would the remaining four among themselves.

*When any reliability subject's (N=50) dual ratings produced unequal scores with a given scale, one was randomly chosen and used in this and subsequent analyses.

#The use in behavior rating scale development of the multitrait-multimethod matrix proposed by Campbell & Fiske (1959) has been endorsed by Johnson & Bolstad (1973).

Criterion Related Validity

Analysis

The five subject groups vary along the dimensions of age, hospital versus community resisdence, and type of illness or reason for hospitalization. It was possible, by appropriately clustering the groups, to measure the scales' sensitivities to each of these dimensions (i.e., concurrent criteria) singly. To this end, four planned comparisons were carried out, using the same data as in the construct validity section of this study. In other words, the relevance of each concurrent criterion (according to which subject groups were themselves grouped) to competence in activities of daily living, as measured by each reliable scale, was assessed, as well as the relative usefulness of each reliable scale in detecting relevant differences between clusters of subject groups. In each comparison, Hotelling's multivariate t-squared test with data from the main scales only was used to determine the presence of significant differences, with alpha set at .05. Findings of significance with this procedure were followed by least significant difference (LSD) multiple comparisons to locate the differences more precisely. Applied in this manner, use of this relatively powerful post-hoc technique does not inflate experiment-wise alpha level.

Comparison one. To examine the sensitivity of the scales to age differences (which were here confounded with generational differences (Schaie, 1977)) among neurology patients, Groups I and II were compared with one another on the basis of data from the five main scales. It was hypothesized that older subjects (Group I) would show more disability than younger subjects (Groups II).

This comparison had precedent in the work of Plutchik et al. (1970), who compared the GRS scores of 50 geriatric inpatients with those of 36 randomly drawn non-geriatric inpatients. The hypothesis that age would be associated with greater disability was clearly supported, in two ways. First, 13 GRS items differentiated between geriatric and non-geriatric inpatients, all in the expected direction, at the .05 level or better. It is of interest that the ADL factor of the GRS (Smith, Bright, & McCloskey, 1977) was the most heavily represented factor among these 13 items. Second, the mean GRS scores of a larger sample of 207 geriatric patients and 36 non-geriatric patients showed highly significant age differences, also in the expected direction. The same conclusion can be drawn based on data from the 85 male participants only.

Comparison two. The second comparison examined the scales' sensitivity to types of illness among elderly inpatients. Groups I, III, and IV were compared using data from the five main scales. Based on the work of Dastoor et al. (1975), it was predicted that Group I (organic) subjects would display more disability than would Group IV (functionally impaired psychiatric) subjects. Dastoor et al. (1975) found that, among 80 newly admitted psychogeriatric patients, organically impaired patients received significantly worse scores on both the GRS and the MSBS than did functionally impaired patients. There is apparently no precedent for including a general medical infirmity group in such a comparison, and the great variability of medical disorders which was included in this sample made prediction of their relative ADL competence difficult.

It could be argued that research comparing psychogeriatric patients with and without organic impairment (e.g., Dastoor et al., 1975) is not the same as research comparing elderly neurological patients and elderly functional psychiatric patients (as in the present study). It is probably true, however, that the latter method yields both (a) purer samples of the target populations, and (b) data which are readily generalizable to more integrated populations.

Comparison three. A broader, but less well controlled assessment of the impact of neurological impairment on ADL skills was also performed, by comparing Groups I and II (elderly and non-elderly neurological subjects) against Groups III through V, using data from the five main scales. It must be recognized, however, that age, in addition to organicity, was unevenly distributed across the comparison, and the latter cluster (Groups III through V) was highly heterogeneous. Naturally, it was to be predicted that the neurologically impaired subjects would fare worse than the otherwise impaired and the healthy subjects.

Comparison four. Finally, Groups I, III, and IV together were compared with Group V to assess the scales' sensitivity to inpatient versus non-patient status among elderly men. The direction of prediction clearly favored the latter group. This analysis was similar to that reported by Lawton (1971), which found different mean PSMS scores between samples of old age home applicants, psychogeriatric admissions, and protective custody patients. The present analysis additionally provided tests of significance.

Expectations of Scale Performance

The capacity of these scales to yield data which accurately reflect differences along meaningful subject variables constitutes an

important dimension along which the scales can be ordered and evaluated. The relative discriminative power of the scales can be evaluated on the basis of the results of each of the above comparisons. When other factors (e.g., reliability, length) are approximately equal, greater discriminative ability would warrant one scale being recommended over another. The very narrow range of both previous reliability estimates (.87 to .94), and present reliability estimates (.85 to .98), and the similarity of scale content, make predictions of the scales' relative strengths very difficult. It was reasonable to expect, however, that those scales which had reportedly discriminated between certain subject categories in the past would replicate these discriminations between comparable categories in the present study. Furthermore, the fact that a given scale had already successfully discriminated between certain groups once made it plausible to predict that the scale would, in fact, discriminate between those same categories in this study <u>better</u> than the other, unproven scales.

Therefore, since the GRS was the only scale in this study to have already once discriminated between elderly and nonelderly mental patients (Plutchik et al., 1970), it was predicted that the GRS would yield the most discriminative data between Groups I and II (Comparison One). Similarly, the GRS and the MSBS, having already been found to differentiate between organically and functionally impaired psychogeriatric patients (Dastoor et al., 1975), were predicted to be the most effective discriminators between subject Groups I, III, and IV (Comparison Two). The existence of rough norms on the PSMS (Lawton, 1971) for groups which probably differ significantly along the

dimension of organic impairment suggested that this scale would trail the GRS and MSBS in discriminative power, but exceed the remaining scales. Finally, with the extension of the organic/nonorganic dimension across ages in Comparison Three, the same predictions of scale effectiveness would hold as in Comparison Two.

CHAPTER III
RESULTS

Interrater Reliability

As can be seen in Table One, all main and additional measures proved reliable at a probability level of .0001, and all but one (GRS 2) well exceeded the cutoff for further study of r=.8. Of the five main scales (RDRS, PSMS, GRS 3, MSBS, & PADL), the PADL proved the most reliable (.985), followed by MSBS (.963), GRS 3 (.924), PSMS (.895), and RDRS (.853).

Construct Validity

Table Two shows the group means and standard deviations (SD) for the entire sample (N=140). Close examination reveals considerable variability in the SD across samples with several tests, especially the PADL. In most cases, this appears attributable to a ceiling effect, whereby many subjects passed all items, thus limiting the variance. The possibility is thus raised that the present data set may violate the assumption of later used multivariate statistical procedures that the sample variance-covariance matrices are equal. Sophisticated statistical programs to compare these matrices do exist, but were not run with these data. The assumption of equality is generally considered to be quite robust, and the effect or meaning of its violation in a multivariate context is not well understood.*

*See footnote p. 36.

TABLE ONE

INTERRATER RELIABILITY

	SCALE	r-VALUE[a]	PROBABILITY
MAIN SCALES	RDRS	.853	.0001
	PSMS	.895	.0001
	GRS3	.924	.0001
	MSBS	.963	.0001
	PADL	.985	.0001
ADDITIONAL SCALES	PSMSX	.873	.0001
	GRS1	.872	.0001
	GRS2	.733	.0001
	GRS_{1+2+3}	.924	.0001
	MSQ	.988	.0001

[a]Pearson Product Moment Correlation.

TABLE TWO

GROUP MEANS AND STANDARD DEVIATIONS

		OLD NEUROLOGY N=30	YOUNG NEUROLOGY N=30	MEDICINE N=30	PSYCHIATRY N=20	COMMUNITY N=30
MAIN SCALES	RDRS	27.933	23.167	23.833	21.600	19.967
	SD	8.982	6.148	6.000	2.186	3.855
	PSMS	2.233	3.667	3.500	5.400	5.567
	SD	2.459	2.482	2.271	0.940	1.073
	GRS3	5.867	2.667	2.500	1.200	1.000
	SD	4.718	3.377	2.980	1.824	1.722
	MSBS	19.633	21.700	21.433	23.350	23.800
	SD	6.457	4.187	4.023	1.872	1.095
	PADL	11.033	13.533	13.600	15.550	15.600
	SD	5.372	3.730	3.500	1.146	0.563
ADDITIONAL SCALES	PSMSX	20.333	25.467	25.433	29.200	29.300
	SD	8.384	5.847	5.444	1.240	1.705
	GRS1	10.900	6.367	8.933	6.400	2.167
	SD	6.205	4.484	5.445	3.761	2.493
	GRS2	0.633	0.133	0.167	0.200	0.067
	SD	1.377	0.434	0.648	0.410	0.254
	MSQ	5.733	8.200	8.200	9.200	9.667
	SD	4.291	2.905	1.901	1.105	0.547
	AGE	69.367	48.833	69.633	63.250	72.167
	SD	8.652	12.149	8.920	3.492	7.697
	NON-WHITE%	16.7	23.3	16.7	0	0

Nevertheless, possible limitations on the interpretability of statistical results obtained cannot yet be ruled out.

Table Three shows the intercorrelations of all main scales and additional measures used. Every correlation is in the expected direction, with increased impairment on one scale reflecting increased impairment on all others, regardless of scoring direction (PSMS, MSBS, & PADL data correlated positively with ability; RDRS & GRS 3 data correlate positively with impairment). Every correlation between reliable scales is significant at p=.0001. The correlations range from .498 to .893.

Table Four shows the results of a principal components analysis of the main scales. Using the customary requirement of eigenvalue equal to or greater than one, the analysis yielded only one interpretable factor, which accounts for 79.1% of the overall variance, and therefore should clearly be labeled ADL. The scale loadings on this ADL factor can be interpreted as indicators of each scale's overall agreement with the other scales. Furthermore, these loadings can be used to rank the scales according to the extent of agreement. GRS 3 ranks first in this regard, with a factor loading of .942, followed by RDRS (.928), PADL (.907), PSMS (.868), and MSBS (.793). The relatively weak agreement of the MSBS with other scales was expected (see METHOD), and is treated in more detail below (see DISCUSSION).

Criterion-Related Validity

Comparison One

The results of this comparison are shown in Table Five. Every scale reveals greater impairment among elderly than non-elderly neurology inpatients. Taken together, however, the five main scales narrowly failed to discriminate between elderly and non-elderly

TABLE THREE

CONVERGENT CONSTRUCT VALIDITY
CORRELATION MATRIX
N = 140

		RDRS	PSMS	GRS3	MSBS	PADL	PSMSX	GRS1	GRS2
MAIN SCALES	RDRS								
	PSMS	.824[a]							
	GRS3	.893	.847[a]						
	MSBS	.624[a]	.498	.630[a]					
	PADL	.761[a]	.670	.791[a]	.814				
ADDITIONAL SCALES	PSMSX	.862[a]	.911	.899[a]	.614	.776			
	GRS1	.715	.674[a]	.731	.624[a]	.622[a]	.671[a]		
	GRS2	.423	.312[□a]	.376	.202[*a]	.407[a]	.344[a]	.357	
	MSQ	.731[a]	.652	.798[a]	.668	.752	.722	.716[a]	.357[a]

All correlations significant at $P \leq .0001$ except:

□ $P \leq .0002$ and

* $P \leq .0169$.

[a] Correlation has been refelcted to eliminate minus sign.

TABLE FOUR

PRINCIPAL COMPONENTS ANALYSIS OF MAIN SCALES

FACTOR	EIGENVALUE	PCT OF VAR
1	3.95426	79.1
2	0.62892	12.6
3	0.16588	3.3
4	0.15193	3.0
5	0.09900	2.0

COMPONENT LOADINGS

FACTOR 1

RDRSN[a]	0.92811
PSMS	0.86849
GRS3N[a]	0.94167
MSBS	0.79335
PADL	0.90689

[a]All data from this scale reflected to remove minus signs.

TABLE FIVE

CRITERION RELATED VALIDITY: COMPARISON ONE

MULTIVARIATE RESULTS[a]

F	PROBABILITY
2.32	.0549

UNIVARIATE RESULTS

		F	PROBABILITY
MAIN SCALES	RDRS	5.75	.0197
	PSMS	5.05	.0285
	GRS3	9.13	.0037
	MSBS	2.16	.1467
	PADL	4.38	.0407
ADDITIONAL SCALES	PSMSX	7.57	.0079
	MSQ	6.80	.0116
	GRS1	10.52	.0020
	GRS2	3.60	.0628

[a]Variables: RDRS, PSMS, GRS3, MSBS, & PADL

neurology inpatients at alpha=.05 (multivariate probability=.0549). Examination of univariate results reveals that this failure was due to the MSBS, which yielded group differences with a random probability of .1467. The other four scales each discriminated significantly between the two age groups.

Such a phenomenon (i.e., multivariate insignificance despite strong univariate results) is neither unusual nor discrepant. At least two factors may have contributed to it occurring in this case. First, a slightly larger sample would have yielded more degrees of freedom, and, therefore, probably, multivariate significance. Second, because of the especially high intercorrelations of the four "pure" ADL scales (see Table Three), they actually accounted for little more total variance than just one of them alone would have. The inclusion of all four, however, nevertheless cost degrees of freedom in the multivariate analysis, decreasing its power. In other words, the inclusion of highly redundant measures created a situation of relatively high covariance and low variance, leading predictably toward multivariate insignificance.*

Comparison Two

As Table Six shows, highly significant overall differences were found between elderly inpatients on three hospital services

*The following individuals were consulted with regard to this and other statistical issues in this study: John Overall, Ph.D., Department of Psychiatry and Behavioral Sciences, University of Texas Medical Center, Houston, Texas; Jerry Lester, Ph.D., Institute for Computor Services and Applications, Rice University, Houston, Texas; Ron Marks, Ph.D., Department of Biostatistics, University of Florida, Gainesville, Florida.

TABLE SIX

CRITERION RELATED VALIDITY: COMPARISON TWO

MULTIVARIATE RESULTS[a]

F	PROBABILITY
4.58	.0001

UNIVARIATE RESULTS

		F	PROBABILITY
MAIN SCALES	RDRS	5.84	.0043
	PSMS	13.56	.0001
	GRS3	12.09	.0001
	MSBS	3.70	.0293
	PADL	8.09	.0006
ADDITIONAL SCALES	PSMSX	13.00	.0001
	MSQ	9.66	.0002
	GRS1	4.18	.0190
	GRS2	2.12	.1274

[a]Variables: RDRS, PSMS, GRS3, MSBS, & PADL

(neurology, medicine, and psychiatry) using data from the five main scales together and from each scale individually. With each scale, neurological patients proved more imparied than psychiatric patients. In every case, furthermore, the mean impairment level of medical patients fell somewhere between neurological and psychiatric means. Table Seven shows the results of LSD multiple comparisons performed on the group means yielded by each scale. In general, the significant overall statistics (Table Six) appear to reflect the gap between the high impairment level of the elderly neurology paitents on the one hand, and the lower impairment level of the medical and psychiatric patients on the other.

Comparison Three

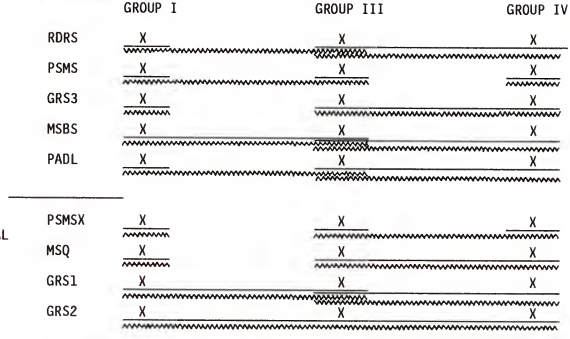
Highly significant differences between neurologically impaired subjects and subjects without neurological impairment were found using data from all main scales together and singly, as shown in Table Eight. Neurology patients showed significantly more deficit than the other groups. In order to determine the sources of these findings of overall differences, another LSD was performed, this time with the data from each scale on the entire sample (N=140), the results of which are shown in Table Nine. The main difference to emerge appears to lie between elderly neurology patients on the one hand, and psychiatric patients and nonhospitalized subjects on the other. With every scale, non-elderly neurology inpatients and elderly medical inpatients fall between these extremes and do not differ significantly from one another.

Comparison Four

Highly significant differences between hospitalized and non-hospitalized subjects were found using all main scales together and

TABLE SEVEN

A POSTERIORI MULTIPLE COMPARISONS OF GROUPS I, III, AND IV[a]



[a] Lines connect groups whose means do not differ significantly by LSD at alpha = .05 (straight lines) and at alpha = .01 (wavy lines).

TABLE EIGHT

CRITERION RELATED VALIDITY: COMPARISON THREE

MULTIVARIATE RESULTS[a]

F	PROBABILITY
5.97	.0001

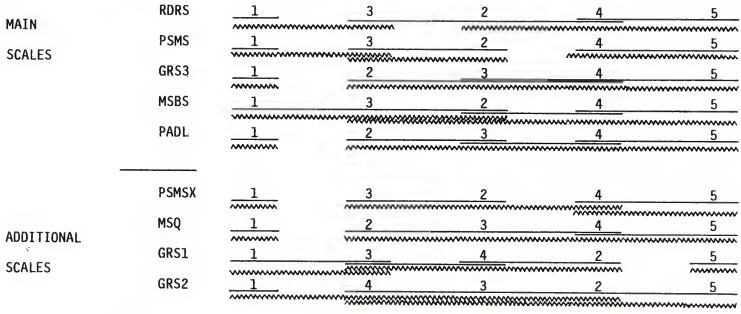
UNIVARIATE RESULTS

		F	PROBABILITY
MAIN SCALES	RDRS	11.79	.0008
	PSMS	23.23	.0001
	GRS3	21.17	.0001
	MSBS	8.81	.0035
	PADL	17.18	.0001
ADDITIONAL SCALES	PSMSX	24.56	.0001
	MSQ	18.81	.0001
	GRS1	9.71	.0022
	GRS2	3.50	.0634

[a]Variables: RDRS, RSMS, GRS3, MSBS, & PADL

TABLE NINE

A POSTERIORI MULTIPLE COMPARISONS OF ALL SUBJECT SAMPLES[a,b]



KEY:

[a] 1 = Elderly Neurology, 2 = Non-Elderly Neurology, 3 = Medicine, 4 = Psychiatry, 5 = Non-inpatient

[b] Lines connect groups whose means do not differ significantly by LSD at alpha = .05 (straight lines) and at alpha = .01 (wavy lines).

alone (Table Ten). In every case, the direction of the comparison was for inpatients to show more impairment than non-hospitalized subjects. Again using the full-sample LSD results shown in Table Nine, the significant differences obtained seem to exist primarily between the non-hospitalized subjects on the one hand and the neurology (especially the elderly) and medicine patients on the other. None of the five scales yielded significant differences between the psychiatric sample mean and the non-hospitalized sample mean.

Relative Discriminative Power of the Scales

In Comparison One, the GRS 3 discriminated most powerfully ($F=9.13$, $p=.0037$), followed by the RDRS ($F=5.75$, $p=.0197$), the PSMS ($F=5.05$, $p=.0285$), and the PADL ($F=4.38$, $p=.0407$). Only the MSBS ($F=2.16$, $p=.1467$) failed to discriminate significantly between the old and young neurology patients.

In Comparison Two, of the five main scales, only the PSMS yielded data which distinguished significantly between medical and psychiatric patients, and in fact maintained this significant difference at alpha=.01. The PSMS thus appears to be the most powerful discriminator between these groups. GRS 3 is the next most powerful because it was the only other scale which maintained a significant difference between two adjacent means at alpha=.01. RDRS and PADL, tied for third, failed to do so, and last-ranked MSBS distinguished significantly only between extreme means (neurology and psychiatry samples) even at alpha=.05.

In Comparison Three, as in Comparison Two, the PSMS proved the most powerful discriminator, in terms of univariate F-value, although the difference between the PSMS, GRS 3, and PADL may be psychologically trivial, given that they all yielded results significant at $p=.0001$.

TABLE TEN

CRITERION RELATED VALIDITY: COMPARISON FOUR

MULTIVARIATE RESULTS[a]

F	PROBABILITY
4.49	.0010

UNIVARIATE RESULTS

		F	PROBABILITY
MAIN SCALES	RDRS	12.48	.0006
	PSMS	20.32	.0001
	GRS3	10.33	.0017
	MSBS	7.95	.0057
	PADL	9.74	.0023
ADDITIONAL SCALES	PSMSX	13.78	.0003
	MSQ	12.92	.0005
	GRS1	41.68	.0001
	GRS2	2.49	.1177

[a]Variables: RDRS, PSMS, GRS3, MSBS, & PADL

The weakness of the RDRS and MSBS is only relative in this case, as they, too, yielded clearly significant results.

In Comparison Four, the scales can be ranked according to discriminative power, as expressed in F-values, as follows: PSMS ($F=20.32$, $p=.0001$), RDRS ($F=12.48$, $p=.0006$), GRS 3 ($F=10.33$, $p=.0017$), PADL ($F=9.74$, $p=.0023$), and MSBS ($F=7.95$, $p=.0057$).

Table Eleven shows the overall hit rate for each scale in distinguishing between elderly neurological patients and elderly psychiatric patients. For each scale, an optimal cutting score was identified, and a Bayseian analysis conducted. The highest hit rate was demonstrated by the PADL (80.85%), followed by the PSMS (78.4%), GRS 3 (77.5%), RDRS (73.3%), and MSBS (65.0%).

TABLE ELEVEN

CLASSIFICATION HIT RATES: GROUPS I AND IV

						HIT RATE (%)	OVERALL* HIT RATE (%)
MAIN SCALES	RDRS	through	25	=	PSYCH	100	
			26+	=	NEURO	46.7	73.3
	PSMS	through	4	=	NEURO	66.7	
			5+	=	PSYCH	90	78.4
	GRS3	through	3	=	PSYCH	95	
			4+	=	NEURO	60	77.5
	MSBS	through	22	=	NEURO	50	
		or	23+	=	PSYCH	80	65.0
		through	23	=	NEURO	70	
			24+	=	PSYCH	60	65.0
	PADL	through	14	=	NEURO	66.7	
		or	15+	=	PSYCH	95	80.85
		through	15	=	NEURO	86.7	
			16+	=	PSYCH	75	80.85
ADDITIONAL SCALES	PSMSX	through	25	=	NEURO	60	
			26+	=	PSYCH	100	80.00
	MSQ	through	5	=	NEURO	46.7	
			6+	=	PSYCH	100	73.35

*Adjusted for unequal sample sizes: NEURO N=30, PSYCH N=20.

CHAPTER IV
DISCUSSION

Review of Findings

This section will first summarize the findings of the present study concerning the interrater reliability, and construct and criterion-related validities of the five main scales. A treatment of each scale in light of these findings will then follow a discussion of psychometric issues of relevance.

The interrater reliability of the main scales proved uniformly high, with even the lowest realibility (MSBS, r=.853) still falling in a range that is quite acceptable for research or assessment purposes. These results closely approximate the interrater reliability estimates obtained by the scales' authors. The repeated demonstration of these scales' interrater reliability suggests that data generated by any single rater will agree substantially with data generated by other raters. It should be cautioned, however, that this assumption may only be a safe one if all raters who use the scale are trained in a manner and with a rigor comparable to that given research raters. Still, in general, the interrater reliability results obtained here constitute, by replicating earlier estimates, a strong basis from which validation work can be justifiably pursued.

The construct validity data provide convincing evidence that these scales, in particular the RDRS, PSMS, GRS 3, and PADL, indeed

measure the same domain of behavior. The large portion of total variance explained by the ADL principal component (79%) is especially impressive. Although the MSBS also correlates with these scales to a high degree of significance, the data confirm the prediction that the MSBS would have lower intercorrelations than the other scales. This is presumably due to the focus of the MSBS on social behavior rather than basic self-care skills. The significant agreement obtained, however, strongly suggests an interactive maintenance or decline of those aspects of these two behavioral domains (social and ADL) which are actually distinct from one another. It is certainly plausible, in other words, that maintaining oneself day to day often depends heavily upon, as well as contributes materially to, maintaining smooth social functioning.

The results of Comparison One indicate that, as hypothesized, older neurological patients show significantly greater impairment of ADL capacity than do younger patients. However, social behavior data reveal only a nonsignificant trend for older patients to be more impaired than younger patients. Therefore, advanced age or the particular neuropathological conditions that tend to accompany it appear to have a much clearer deleterious effect on these patients' basic self-care skills than on their social behavior.

Comparison Two was a replication and extension of previous comparisons of ADL skill levels of neurologically and functionally impaired elderly inpatients. The GRS 3 and MSBS indeed replicated their earlier discriminations between these groups, as was hypothesized. Every other scale did, too, however. The medical sample consistently ranked between the neurological and functional groups,

indicating that among the elderly, neurological pathology tends to be more devastating to basic functioning than does visceral pathology without primary neurological involvement. The inability of all scales except the PSMS to discriminate significantly between medical and psychiatric samples indicates that the PSMS can be more sensitive than the other scales to fine differences in ADL skill level.

As hypothesized, the results of Comparison Three show that, even when the elderly neurology patients are clustered with their generally less impaired, younger counterparts, the effect of neuropathology is still, as found in Comparison Two, greater impairment of ADL and basic social skills than is found among subjects without neuropathology. As mentioned, this is a poorly controlled contrast, partially confounding hospital status with neurological status. It is of note that no main scale could draw a significant distinction between the psychiatric and nonhospitalized samples, which together formed the primary contrast to the highly impaired elderly neurology patients.

The prediction that hospitalized elderly, as a group, would show greater impairment than non-hospitalized elderly of the behavioral skills measured was confirmed by the outcome of Comparison Four. This appears to hold equally for ADL and basic social skills. Again, these significant differences emerged despite the essentially normal profile of the psychiatric sample.

In sum, with one exception (MSBS is Comparison One), these results allow every scale evaluated to claim substantial interrater reliability and construct validity. Also, criterion-related validity for several populations has been demonstrated along the following dimensions: (1) age among male neurology inpatients, (2) type of medical impairment

among elderly male inpatients, within broad catagories, (3) presence of neurological pathology among males, and (4) inpatient status among elderly males.

Limitations and Biases of the Study

In evaluating the internal validity of this study, three factors which proved especially problematic require discussion. These are (1) recruitment procedures, (2) selection and preparation of raters for subject-absent scales, and (3) race of subjects. The primary limitation on this study's external validity will then be discussed.

(1) Recruitment of inpatients suffered perhaps most severely from variability in the promptness and diligence with which eligible patients were approached for participation. Although exhaustive and clear data on this do not exist, it is likely that many eligible patients were lost as participants due to tardy recuitment or testing. This was particularly evident on the medical wards, where patient turnover is relatively quick. If these patients with shorter hospitalizations were thus less likely to participate, the data might reflect a bias toward greater impairment.

Among inpatients, medical patients were also the most likely to refuse participation for lack of interest. These refusers were often noted, anecdotally, to be fairly spry, intact, and sometimes cynical individuals who probably would have fared well on the scales. Differential subject interest in the project may also, therefore, have lent the data a bias toward greater impairment of medical subjects. This bias risk is moderated somewhat by the fact that a few uninterested refusers did so because of significant pain. One such patient died two days after refusing.

Recruitment of nonhospitalized subjects posed possibly the most serious methodological difficulty of the entire study. In contrast to the inpatient recruitment baises discussed above, it is probable that the nonhospitalized sample data are significantly biased toward more intact ADL skills. Since all 140 subjects were administered the MSBS and PADL at GVAH, no housebound community residents participated. Since early plans to recruit nonhospitalized veterans through veterans organizations such as the Veterans of Foreign Wars (VFW) proved totally untenable, a variety of more accessible subject pools were tapped. The largest portion of the final sample was recruited from among the retired men who do volunteer work at GVAH. Their voluntary duties consist primarily of escorting patients and carrying hospital paperwork between the various wards and clinics in GVAH. These men are uniformly alert, ambulatory, and fit. A large group of less healthy participants was recruited from among veterans making outpatients visits to the GVAH evaluation or specialty clinics. The uncomfortably arbitrary, rather than systematic, nature of the sampling procedure used to recruit nonhospitalized subjects makes it very difficult to call the sample data representative of elderly nonhospitalized male veterans in general. It is strongly recommended that future studies comparing these scales across the dimension of hospitalization seriously attempt, within practical limits, to obtain representative data through systematic sampling. This will probably require the cooperation of the VA itself, or of viable local units of the VFW or American Legion, and the use of their membership roles.

(2) Less than ideal circumstances also prevailed in the utilization of raters for the three subject-absent scales. Except in rare

cases of unusual rater availability, two or three of the scales were often completed by the same rater for each subject in any group. It is possible, as a result, that the intercorrelations of these three scales (see Table Three) are spuriously high. The risk of such a rater effect contaminating the construct validity results is probably limited by the excellent interrater reliability of these scales. Nevertheless, the possibility of such contamination cannot be ruled out without statistically comparing the data yielded by various raters. However, the distribution across subject groups of the various raters, both as individuals and as groups (nurses, residents, medical students, etc.), was unplanned and based largely on immediate exigencies, and therefore extremely unbalanced. This would unfortunately tend to make such statistical comparisons quite difficult.

An additional source of poorly controlled variance is probably present in the subject-absent data on community subjects. Whereas all raters of inpatients were professional or paraprofessional hospital staff members and received personal orientations and instructions in the purpose and use of the scales, this was not always the case with raters of nonhospitalized subjects. With perhaps half of these subjects, the scales were explained and handed to the subject, with careful instructions to pass both the scales and the explanation of their use onto their rater (usually spouse or child). The forms were then returned to the researcher in pre-addressed, postage-paid envelopes. Although practically all scales were returned promptly, appropriately completed, and bearing the raters' signatures, the possibilities of confusion and non-independent rating decisions cannot be disconfirmed. Clearly, direct rater training and assurance of

independent rating would have been desirable. Perhaps the most important ramification of the procedure used is the possibility that statistical differences obtained between Group V and the other subject groups cannot be interpreted as reflecting pure subject-related variance. Rather, methodological variance may also be reflected.

(3) As Table Two shows, there is significant variability between samples in the rate of black patient participation in the study. The non-elderly neurology sample was 23.3% black, the elderly neurology and medicine samples were each 16.7% black, and the psychiatry and non-hospitalized samples included no blacks at all. It is meaningful to examine both (a) the comparability of these percentages with system-wide VAH inpatient base rates, and (b) the possible causes and meanings of this inter-sample variability.

Racial percentage information* on VAH inpatients is extracted from two parameters nationwide: bed census on a "typical day," and hospital discharges. These figures combine data from general medical hospitals and long-term care facilities, and are not broken down by age. On a "typical day" (the last Wednesday in September, 1979), out of 68,689 VAH beds nationwide, 77.2% were occupied by whites, and 15.8% by blacks. During fiscal year, 1979, of 903,466 discharges, 77.9% were white, and 16.6% were black. The percentages obtained in this study with medicine and elderly neurology samples are highly

*Source of VA system information: Phone communication, 7/30/80, Dr. William Page, Chief, Division of Biometrics, VA Central Office, Washington, D.C., 202-389-3458.

consistent with these figures, while the non-elderly neurology sample was somewhat above, and the remaining samples significantly below these figures.

During the period of data collection for this study, every elderly black patient on the GVAH psychiatric wards carried a diagnosis of organic impairment, making them ineligible for participation. The all-white psychiatric sample obtained, therefore, is clearly not a product of recruitment or consent bias, but more likely of institutional or sociomedical selection factors such as diagnostic biases, or racial differences in hospital use patterns. There may also be racial differences in early- or late-life exposure to situations or use of substances which compromise central nervous system integrity.

In the community sample, however, the possiblity of an insidious, though unintended, racial bias in the recruitment of subjects is distinctly greater. Although there are few or no blacks among the volunteer force at GVAH, black veterans are surely well represented among clinic outpatients. The unsystematic, somewhat opportunistic recruitment procedure used to select this sample may well have allowed a serious bias against black participation to develop.

The impact of the particular pattern of black participation obtained on the study's test-related conclusions, and on the generalizability of the results to black populations, can be assessed by studying the tests' correlations with race. Out of fifteen correlations (five tests by three groups with black subjects), only one was significant (PSMS, Group I, $r=.4376$, $p=.015$), which is about what might be expected by chance at the stated alpha level of .05. When all 140 subjects are pooled, the greatest test by race correlation is

still nonsignificant (MSBS, r=-.081, p=.338). While certain methodological flaws are apparent, therefore, it still appears that race accounts for very little test variance, and that the present results may be generalizable to black as well as white populations.

Based on the findings of Smith, Bright, and McClosky (1977) that women in a large study showed significantly greater impairment of ADL skills than men, the results of the present study cannot be generalized to female populations. It would be inappropriate to treat the results as valid for female populations.

In addition, it must be recognized that the results of this and earlier studies of these scales cannot be immediately generalized to other populations of male subjects. To attempt such generalizations without first replicating earlier findings with each new population would be cavalier and unscientific.

Choosing a Scale

Despite the limitations and biases discussed above, the results of this study and previous research indicate that these five scales are psychometrically sound in several important respects. They are all highly reliable between raters, they all show high agreement among themselves, and they all (with one exception) proved very sensitive to functional competence variability associated with selected relevant criteria. None failed to replicate earlier reported specific discriminations between groups. Consumers of ADL scales are therefore placed in the enviable position of having to choose the most appropriate measure from among a group of highly qualified instruments. They have the luxury of making choices on the basis of various clinical, pragmatic, and situation-specific factors without sacrificing psychometric quality.

Subject-Absent vs. Subject-Present Scales

One decision a consumer of ADL scales may have to make first in choosing a scale is between subject-absent and subject-present instruments. This decision may have several elements. First, the present study demonstrated that both types can have excellent interrater reliability, criterion-related validity, and correct classification rate. However, the best interrater reliability and lowest misclassification rate were obtained by a subject-present scale (PADL), whereas the subject-absent scales generally dsicriminated best among groups. Second, logistical considerations point out important uniquenesses of the two methods. Use of subject-absent scales assumes rater familiarity with subjects, whereas subject-present measures do not. Familarity, it was assumed in the present study, takes at least a week to develop, and can be an untenable limitation for some assessment purposes. Third, subject-present scales require carefully standardized administration, and thus more extensive rater training and practice than do subject-absent scales. On the other hand, the involvement of fewer personnel may be required when using subject-present scales. Fourth, proper administration of the subject-present scales requires space and special materials, as well as the subject's time and cooperation. Subject-absent scales require none of these, and are more portable. Finally, whereas retesting with subject-absent scales presents no major difficulties, retesting can be quite problematic with subject-present scales, unless a considerable inter-test interval is allowed. Contamination of result generalizability from practice effects on specific test tasks is possible. Also, the requirement that certain standardized administration procedures appear accidental to the subject (MSBS) would be difficult to fulfill in

subsequent administrations.

A synthesis of the above considerations suggests the following guidelines for people faced with the choice between ADL assessment methods. Subject-absent scales are probably the technology of choice (1) whenever many indigenous raters are available and trainable, (2) when there may be lack of subject motivation to be assessed, and (3) especially when retesting of subjects is planned or desirable. This availability of raters, often professional, will exist in most institutions, especially long-term facilities, where familiarity poses no problem. Subject-present scales, on the other hand, seem most appropriate for research and screening purposes involving one-time evaluation of subjects. These scales encourage the repeated involvement of a very few trained raters who need know nothing about the subject except what they learn during testing. The lack or inaccessibility of indigenous, familiar, and trainable raters is no barrier with these scales, but retesting may be.

Summary of Findings on Each Scale

The PSMS, actually the briefest scale studied, proved to be the best of the three subject-absent scales. It was the strongest discriminator of all the scales on three of four validity criteria, and had the best classification hit rate within its method between neurological and psychiatric patients. This scale, as published, is exceedingly simple, with a scoring procedure that ignores more information than it recognizes. Systematic comparisons in this study of the standard scoring procedure (PSMS) with a more comprehensive one (PSMSX, in which proportional scores for partial competence on each item were given, rather than only giving credit for full competence on

each item), however, failed to yield noticable gains in psychometric strength. The brevity, simplicity, and psychometric quality of this scale recommend it highly for applied use and further study. It is clearly the scale of choice when medical patients are to be included in comparisons.

The GRS was described above (see METHOD) as having the strongest base of psychometric support among the scales chosen for study. It was therefore predicted that its ADL factor would prove the best in the present comparisons of psychometric quality. Although this did not prove to be strictly the case (except in Comparison One), GRS 3 clearly emerged as a highly reliable, useful, and powerful scale. As an ADL scale, it is more than adequate. Perhaps the greatest strength of GRS 3 is that it is unlikely to be used in isolation, and thus will provide its users with information on the other GRS factors, Withdrawal/Apathy and Antisocial Disruptive Behavior. Data obtained in the present study indicate that GRS 2 needs a great deal of improvement before it can stand on its own psychometrically, but this limitation is outside the scope of this study, and is only a relative detractor from the considerable overall quality of the GRS.

The RDRS, also, emerged as a quite adequate scale of ADL in all psychometric respects, but was generally the weakest of the ADL scales studied. Its relative merit lies, therefore, less in its present performance than in earlier reported empirical data which (1) demonstrated various criterion-related validities with nursing home patients, and (2) generated norms on a sample of 1000 male nursing home residents.

The PADL demonstrated the best interrater reliability and the lowest misclassification rate between old neurological and psychiatric patients of all scales studied, but trailed the subject-absent scales slightly on criterion-related validity tests. Its strengths coincide with those of subject-present scales described in general above--it can save time in certain situations and make measurement possible when appropriate raters are not accessible or do not coexist naturally with subjects. Although special equipment is needed, it is not expensive, and administration is very quick.

Although the MSBS was clearly the weakest of the five scales with regard to the criterion-related comparisons and hit rate calculations performed in this study, it must be recognized as a respectable measure in its own right which failed to yield significant results in only one comparison. The relative weakness of its convergent validity with obvious measures of ADL is consistent with its intended construct content, and points out only the need for studies comparing it directly with other measures of minimal social behavior. It appears that this construct may be less sensitive to health status, the primary focus of health care delivery systems, than is ADL. As mentioned above (p. 18), the MSBS contains the Mental Status Questionnarie, a ten-item index of orientation to time and place, and of memory for typically overlearned facts. Despite the significant similarities between MSBS data and ADL scale data, the practicality of the MSBS is undermined by the fact that the much briefer MSQ excels the MSBS on every test in this study.

Further Research

The most pressing research needs with regard to the five scales evaluated here are no longer psychometric, but rather research aimed at testing and improving the scales' applicability to clinical tasks, such as individual assessment and treatment evaluation. Perhaps the most basic need is for good normative data for the scales on a wide variety of subject populations. Such normative data would allow the generalizability of research findings using the scales to be properly assessed. Representative sampling procedures should be pursued with much energy in generating these normative samples. The availability of normative distribution data will allow the placement of individual subjects along a realistic continuum of competence.

Practical considerations such as subject availability contributed to the present study being limited to men only. Much previous work with these scales (e.g., Linn, et al., 1977), also, involved only males. Previous findings of sex differences in ADL skill among the elderly (Smith, Bright, & McClosky, 1977) make it important to control for sex in this research area. Needless to say, however, the need for quality assessment techniques is equally great with female populations. New norms should include equal but separate female samples. Validity studies already completed with male subjects should be replicated with female subjects.

Longitudinal research to investigate the relative stability/sensitivity to change of various ADL scales should also be conducted. Such assessments of test-retest reliability would most informatively be done using several ADL scales in the same study, so that, as in the present study, the data will be directly comparable. This is the best

approach to the dilemma of identifying true, underlying stability or change. One clinically useful context in which such data might be gathered would be a long-term project plotting the clinical course of a sample of patients with a selected degenerative neurological disease.

ADL skill is the behavior domain targeted by Reality Orientation Therapy. As amply demonstrated above, the research to date has not appropriately evaluated the effectiveness of this treatment. This study has demonstrated that appropriate and high quality instruments are available for a proper evaluation of ROT and any other treatment designed to affect ADL competence. This research should now proceed.

The relationship of ADL competence to performance on popular intellectual and neuropsychological measures could be studied with relative ease, simply by adding a sound ADL instrument to the neuropsychological battery. It is reasonable to predict that ADL competence would be found to behave in general more like other measures of old learning (Information, Vocabulary, Comprehension) than like measures of new learning and higher level adaptive functioning (Digit Symbol, Block Design, Form Board). This is because, in most neurological conditions frequent among elderly patients, concrete overlearned behaviors generally prove more robust and tenacious than behaviors which are rarely elicitted or only selectively appropriate, requiring complex cognitive mediation.

The classification hit rates of the ADL scales studied here (73.3% - 80.85%) suggest that the scales may be valuable additions to the battery used by clinical psychologists frequently asked to assist in a differential diagnosis between depression and organic impairment

in elderly patients. Of course, the cutting scores used in this study to calculate hit rates should not be directly applied to new samples or other populations. Widely applicable cutting scores must be based on extensive normative data. Independent validation is also required because an upward bias is inherent when any statistic is both calculated from and tested on the same sample. Nevertheless, studies assessing the incremental validity of ADL scales in such contexts as depression versus organicity decisions would be valuable.

Significance of the Study

With the information generated by this study, future investigators researching the behavioral outcome of geriatric treatment programs are now better able to make objective and defensible choices of dependent variable measures. As explained above, past difficulties involved in making such choices has had a noticeably deleterious effect on the quality of some gerontological research. This study is a positive forward step in the crucial, yet understudied, areas of scale reliability and validity. Future gerontological researchers who heed and use the results of this study will obtain, in their own work, findings which are more accurate and more credible, and therefore of much more use to the administrative planners who must ultimately apply those findings. The clinician's individual assessment armamentarium has also been augmented by this study. Thus, this project offers important contributions to the development of geriatric care delivery systems on the clinical, research, and administrative levels.

APPENDIX A

SUBJECT INFORMED CONSENT FORM

PARTICIPANT'S CONSENT FORM

PROJECT TITLE: VALIDATION STUDY OF SEVERAL GERIATRIC RATING SCALES.
PRINCIPAL INVESTIGATORS: PAUL K. CHAFETZ & DR. W. RICE

NAME _____ ADDRESS _____

HOSPITAL No. (IF PATIENT) _____ DATE_____ TIME _____ WARD_____

I agree to participate in the research described below:

This is a study comparing five behavior rating scales often used to measure people's activities in areas such as eating, grooming, socializing, and moving about. With information from participants such as yourself, we hope to choose the best of these scales for further use. Although the knowledge gained by this study may not be of direct or immediate benefit to you, it will be very useful in the evaluation of new treatments for people like you.

As a participant, you will be rated with all five scales. The nurses will complete three of the scales without you during their rounds. That is, they will complete a brief questionnaire, based on their informal observations of your behavior. The other two scales will be completed with you personally by a project investigator, and will take less than one hour in all. You will be asked to answer some short questions, and to perform a few daily behaviors, such as using a spoon. Participation in this study does not involve any additional medical procedures, so the only discomfort you might experience is slight fatigue during testing. Information gained through these scales may be shared with your treatment team and with the staff of this research project, but is otherwise strictly confidential.

Your participation is very much appreciated, and we will gladly answer, as well as we can, any questions you may have.

I have read, and understand, the nature and purpose of this research, as described above. Furthermore, it is agreed that the information gained from this investigation may be used for educational purposes which may include publication. I understand that I may withdraw my consent, and discontinue my participation, at any time without prejudice. I also understand that, in the event of my sustaining a physical injury which is proximately caused by this experiment, no professional medical care will be provided me without charge.

WITNESS_____ SIGNED _____

I, the undersigned, have defined and fully explained the nature and purpose of this project to the above person and/or the person authorized to consent for him.

SIGNED _____

APPENDIX B

SCALES

WARD _____ Scale 1 (RDRS) Date _____

Subject's Name _____ Rater's Name _____

Directions: On the basis of your knowledge about the subject, at the present time, will you please rate the following items:

EATING

__ no assistance
__ moderate assistance
__ considerable assistance

DIET

__ regular diet
__ modified regular diet
__ special diet

REQUIRES MEDICATION

__ rarely
__ occasionally
__ every day

SPEECH

__ not impaired
__ moderately impaired
__ unable to be understood

HEARING

__ normal
__ moderately impaired
__ deaf

SIGHT

__ normal (glasses)
__ moderately impaired
__ blind

WALKING

__ no assistance
__ crutches; someone's help
__ unable to walk

BATHING

__ no assistance
__ moderate assistance
__ considerable assistance

DRESSING

__ no assistance
__ moderate assistance
__ considerable assistance

INCONTINENCE

__ never
__ occassionally
__ all the time

SHAVING

__ no assistance
__ moderate assistance
__ considerable assistance

SAFETY SUPERVISION

__ never
__ sometimes
__ all the time

CONFINED TO BED

__ not at all
__ part of the day
__ all the time

MENTALLY CONFUSED

__ never
__ occasionally
__ all the time

UNCOOPERATIVE

__ never
__ occasionally
__ all the time

DEPRESSION

__ never
__ occasionally
__ all the time

WARD _____ Scale 2 (PSMS) DATE_____

Subject's Name _____ Rater's Name _____

TOILET **Check one answer in each category.

__ Cares for self at toilet completely; no incontinence
__ Needs to be reminded, or needs help in cleaning self, or has rare accidents (weekly at most)
__ Soiling or wetting while asleep, more than once a week
__ Soiling or wetting while awake, more than once a week
__ No control of bowels or bladder

FEEDING

__ Eats without assistance
__ Eats with minor assistance at meal times, with help in preparing food or with help in cleaning up after meals
__ Feeds self with moderate assistance and is untidy
__ Requires extensive assistance for all meals
__ Does not feed self at all and resists efforts of others to feed him

DRESSING

__ Dresses, undresses, and selects clothes from own wardrobe
__ Dresses, and undresses self, with minor assistance
__ Needs moderate assistance in dressing or selection of clothes
__ Needs major assistance in dressing but cooperates with efforts of others to help
__ Completely unable to dress self and resists efforts of others to help

GROOMING (Neatness, hair, nails, hands, face, clothing)

__ Always neatly dressed and well-groomed, without assistance
__ Grooms self adequately, with occasionally minor assistance, e.g., in shaving
__ Needs moderate and regular assistance or supervision in grooming
__ Needs total grooming care, but can remain well groomed after help from other
__ Actively negates all efforts of others to maintain grooming

PHYSICAL AMBULATION

__ Goes about grounds or city
__ Ambulates within residence or about one block distance
__ Ambulates with assistance of (check one): (__a) another person, (__b) railing, (__c) cane, (__d) walker, or (__e) wheelchair: (__1) gets in & out without help (__2) needs help getting in & out
__ Sits unsupported in chair or wheelchair, but cannot propel self without help
__ Bedridden more than half the time

(PSMS, p. 2)

BATHING

__ Bathes self (tub, shower, sponge bath) without help
__ Bathes self, with help in getting in and out of tub
__ Washes face and hands only, but cannot bathe rest of body
__ Does not wash self, but is cooperative with those who bathe him
__ Does not try to wash self, and resists efforts to keep him clean

Scale 3
(GRS)

WARD _____ DATE _____

Subject's Name _____ Rater's Name _____

Directions: Check one answer under each question.

WHEN EATING, THE SUBJECT REQUIRES:

__ No assistance (feeds himself)
__ A little assistance (needs encouragement)
__ Considerable assistance (spoon feeding, etc.)

THE SUBJECT IS INCONTINENT:

__ Never
__ Sometimes (once or twice per week)
__ Often (three times per week or more)

WHEN BATHING OR DRESSING, THE SUBJECT NEEDS:

__ No assistance
__ Some assistance
__ Maximum assistance

THE SUBJECT WILL FALL FROM HIS BED OR CHAIR UNLESS PROTECTED BY SIDE RAILS:

__ Never __ Sometimes __ Often

WITH REGARD TO WALKING, THE SUBJECT:

__ Has no difficulty
__ Needs assistance in walking
__ Does not walk

THE SUBJECT'S VISION, WITH OR WITHOUT GLASSES, IS:

__ Apparently normal
__ Somewhat impaired
__ Extemely poor

THE SUBJECT'S HEARING IS:

__ Apparently normal
__ Somewhat impaired
__ Extremely poor

WITH REGARD TO SLEEP, THE SUBJECT:

__ Sleeps most of the night
__ Is sometimes awake
__ Is often awake

DURING THE DAY, THE SUBJECT SLEEPS:

__ Sometimes __ Often __ Most of day

(GRS, p. 2)

WITH REGARD TO RESTLESS BEHAVIOR AT NIGHT, THE SUBJECT IS:

__ Seldom restless
__ Sometimes restless
__ Often restless

THE SUBJECT'S BEHAVIOR IS WORSE AT NIGHT THAN IN THE DAYTIME:

__ Never __ Sometimes __ Often

WHEN NOT HELPED BY OTHER PEOPLE, THE SUBJECT'S APPEARANCE IS:

__ Almost never sloppy
__ Sometimes sloopy
__ Almost always sloppy

THE SUBJECT MASTURBATES OR EXPOSES HIMSELF PUBLICLY:

__ Never __ Sometimes __ Often

THE SUBJECT IS CONFUSED (UNABLE TO FIND HIS WAY AROUND WARD OR HOUSE, LOSES THINGS):

__ Almost never __ Sometimes __ Often

THE SUBJECT KNOWS THE NAMES OF:

__ More than one member of the staff
__ Only one member of the staff
__ None of the staff

THE SUBJECT COMMUNICATES IN ANY MANNER (BY SPEAKING, WRITING OR GESTURING) WELL ENOUGH TO MAKE HIMSELF EASILY UNDERSTOOD:

__ Almost always __ Sometimes __ Almost never

THE SUBJECT REACTS TO HIS OWN NAME:

__ Almost always __ Sometimes __ Almost never

THE SUBJECT PLAYS GAMES, HAS HOBBIES, ETC.:

__ Often __ Sometimes __ Almost never

THE SUBJECT READS BOOKS OR MAGAZINES:

__ Often __ Sometimes __ Almost never

THE SUBJECT WILL BEGIN CONVERSATIONS WITH OTHERS:

__ Often __ Sometimes __ Almost never

THE SUBJECT IS WILLING TO DO THINGS ASKED OF HIM:

__ Often __ Sometimes __ Almost never

THE SUBJECT HELPS WITH CHORES ON THE WARD:

__ Often __ Sometimes __ Almost never

WITHOUT BEING ASKED, THE SUBJECT PHYSICALLY HELPS OTHER PATIENTS:

__ Often __ Sometimes __ Almost never

(GRS, p. 3)

WITH REGARD TO OTHER PEOPLE ON THE WARD, THE SUBJECT:

__ Has several friends
__ Has just one friend
__ Has no friends

THE SUBJECT TALKS WITH OTHER PEOPLE ON THE WARD:

__ Often __ Sometimes __ Almost never

THE SUBJECT HAS A REGULAR WORK ASSIGNMENT:

__ Away from the ward
__ On the ward
__ No regular assignment

THE SUBJECT IS DESTRUCTIVE OF MATERIALS AROUND HIM (BREAKS FURNITURE, TEARS UP MAGAZINES, ETC.):

__ Never __ Sometimes __ Often

THE SUBJECT DISTURBS OTHER PATIENTS OR STAFF BY SHOUTING OR YELLING:

__ Never __ Sometimes __ Often

THE SUBJECT STEALS FROM OTHER PATIENTS OR STAFF MEMBERS:

__ Never __ Sometimes __ Often

THE SUBJECT VERBALLY THREATENS TO HARM OTHER PATIENTS OR STAFF:

__ Never __ Sometimes __ Often

THE SUBJECT PHYSICALLY TRIES TO HARM OTHER PATIENTS OR STAFF:

__ Never __ Sometimes __ Often

WARD _____ Scale 4 (MSBS) DATE _____

Subject's Name _____ Rater's Name _____

A. E goes to S and introduces himself: "I am _____, Mr. _____. I'm glad to meet you," extending hand

1) Score + if any discernible response to greeting 1) __
2) Score + if response if verbal and appropriate 2) __
3) Score + if S offers hand to E 3) __

B. E says either (a) "Won't you have a seat?" or
(b) "May I sit with you for a while?"
depending on whether S is brought to E, or E comes to S.

4) Score + if (a) S sits without urging or
(b) S assents or acknowledges E's comment 4) __

C. E says, "How are you today?"

5) Score + if any discernible response to question 5) __
6) Score + if response is vergal and appropriate 6) __

D. E drops pencil by pushing it off desk, ostensibly by accident. If S does not pick up pencil spontaneously, E says, "Would you pick up the pencil for me?"

7) Score + if S picks up pencil at all 7) __
8) Score + if S picks up pencil spontaneously 8) __

E. E says "I have something I want to show you." E hold in front of S Figure A of the Bender Gestalt Test.

9) Score + if S looks at Bender card. 9) __

F. "Here is a pencil." E offers it to S, puts paper in front of S, and says, "I would like you to copy this drawing on this paper."

10) Score + if S accepts pencil without further urging 10) __
11) Score + if S makes any mark on paper 11) __
12) Score + if S draws an appropriate circle and 4-sided figure 12) __

G. E says, "How are you getting along?"

13) Score + if any discernible response to the question 13) __
14) Score + if response if verbal and appropriate 14) __

H. E crumples a scap of paper and tosses it at a wastebasket previously placed next to S, purposely missing.

15) Score + S spontaneously picks up paper and deposits it in wastebasket. 15) __

(MSBS, p. 2)

I. E says, "I have a few questions I would like to ask you."

a. Where are we now? (Correct name of place)
b. Where is this place? (Correct city)
c. What is today's date? (Day of month)
d. What month is it?
e. What year is it? MSQ) __
f. How old are you?
g. What is your birthday? (Month)
h. What year were you born?
i. Who is the President of the United States?
j. Who was President before him?

16) Score + if S makes any verbal response, irrespective of content, to all questions 1-10. 16) __

J. E places a magazine in front of S and busies himself with writing on pad while saying, "I'll be busy a minute."

17) Score + if S turns at least one page of magazine. 17) __

K. E rises and extends hand, saying, "Thank you very much, Mr. _____."

18) Score + if S acknowledges E's departure either verbally or with gesture. 18) __

L. The remainder of the items are based on E's judgement of the behavior of the patient throughout the interview:

19) Score + unless inappropriate grimaces or mannerisms are readily apparent 19) __
20) Score + if the patient at any time looks E in the eye 20) __
21) Score + unless S obviously appears to avoid E's gaze at all times, or stares at E fixedly 21) __
22) Score + unless S site in a bizarre position or is in constant motion or is nearly motionless 22) __
23) Score + unless S's clothes are obviously disarranged, unbuttoned, or misbuttoned 23) __
24) Score + unless S is drooling or nasal mucus is visible or food deposits are conspicuous on clothes or face 24) __
25) Score + unless S attempts to move away from E before termination of interview with explanation 25) __

WARD _____ Scale 5 (PADL) DATE _____

Subject's Name _____ Rater's Name _____

SCORING: 0 = does not do task reasonably well on own; requires active personal assistance.
1 = does task reasonably well on own or with some mechanical aid
9 = inapplicable or not asked

"I am going to ask you to do some tasks for me. Some of the things I ask you to do may seem silly or too easy, but I ask everyone to do the same things."

Place paper cup (empty) in front of patient: "Show me how you drink from this cup."

1) Grasps cup in hand 1) __

2) Lifts cup upright 2) __

3) Touches cup to mouth 3) __

4) Tips cup as if drinking 4) __

Place tissue box in front of patient: "Show me how you use a tissue to wipe your nose."

5) Takes tissue out of box 5) __

6) Directs tissue to nose 6) __

7) Makes wiping motions 7) __

Place comb on table in front of patient: "Show me how you comb your hair."

8) Picks up comb 8) __

9) Grasps comb properly 9) __

10) Brings comb to hair 10) __

11) Makes combing motion 11) __

Place emery board in front of patient: "Show me how you file your nails; Use this."

12) Takes emery board in hand 12) __

13) Applies emery board to nail 13) __

14) Files nail 14) __

(PADL, p. 2)

Place shaver in front of patient: "Show me how you use this,"

15) Takes shaver in hand 15) __

16) Brings to appropriate place on face 16) __

17) Makes shaving motions 17) __

Place rubber ball on spoon and place spoon on flat surface in front of patient: "Show me how you lift this spoon with the ball on it to your mouth."

18) Grasps spoon by handle 18) __

19) Keeps spoon horizontal 19) __

20) Keeps ball balanced on spoon 20) __

21) Aims at mouth 21) __

22) Touches ball to mouth 22) __

Place faucet on block in front of patient: "Show me how you use this faucet; Turn it on for water and off again."

23) Puts hand on faucet 23) __

24) Turns faucet in one direction (on) 24) __

25) Turns faucet in opposite direction (off) 25) __

Place light switch in front of patient: "Show me how you turn this switch on for light, and off again."

26) Puts finger on switch 26) __

27) Flicks switch in one direction 27) __

28) Flicks back to original position 28) __

Give patient jacket with long sleeves: "Put this jacket on for me."

29) Takes hold of jacket 29) __

30) Slips one arm in jacket arm 30) __

31) Pulls jacket over shoulders and back 31) __

32) Slips other arm into sleeve 32) __

(PADL, p.3)

"Take it off now."

33) Frees one arm 33) __

34) Frees other arm 34) __

35) Removes jacket from body 35) __

Point to button on jacket. Ask irrespective of ratings above: "Button that for me."

36) Places hand on button 36) __

37) Slips button into button hole 37) __

"Now unbutton it."

38) Places hand on button 38) __

39) Releases button from hole 39) __

Place slipper on floor in front of patient: "Put this slipper on: and take it off."

40) Moves foot to slipper 40) __

41) Inserts foot into slipper 41) __

42) Takes foot out of slipper 42) __

"Do you have real teeth or dentures? Tell me how you take care of your teeth." Place toothbrush in front of patient: "Show me how you brush your teeth. Use this toothbrush."

43) Takes brush in hand 43) __

44) Point bristles toward teeth 44) __

45) Puts brush in or near mouth 45) __

46) Makes brushing motions 46) __

If patient has removable false teeth, ask: "Would you mind taking out your false teeth? Put them back in."

47) Removes teeth 47) __

48) Replaces teeth 48) __

(PADL, p.4)

Place phone in front of patient: "Use this phone to call information (Please dial '0' for Operator)."

49) Picks up receiver 49) __

50) Puts phone to ear 50) __

51) Puts hand on dial 51) __

52) Dials number 52) __

Place paper and pencil in front of patient: "Write your name on this paper."

53) Takes pencil in hand 53) __

54) Moves pencil to paper 54) __

55) Makes writings on paper 55) __

56) Writes name appropriately or legibly 56) __

Show patient doorknob and lock, and lay key down on table; say: "This door is locked. Show me how you open it. Open the lock first. Use this key. Turn the knob and lock it again. Take the key out."

If patient cannot manage lock, rate 0 for items referring to lock. Ask patient: "Show me how you'd open the door if it weren't locked," and rate items for knob.

57) Takes key in hand 57) __

58) Touches lock with key 58) __

59) Inserts key in lock 59) __

60) Turns key in lock (to open) 60) __

61) Grasps doorknob 61) __

62) Turns doorknob 62) __

63) Turns key (to relock) 63) __

64) Takes key out of lock 64) __

Show patient clock face: "What time does this clock say it is?"

65) Reads one hand correctly 65) __

66) Reads other hand correctly 66) __

APPENDIX C

PRACTICAL PROBLEMS OF DOING RESEARCH

This project involved no complex technology or instrumentation. Materials included mimeographed, paper and pencil scales, and a variety of props that were either readily available (e.g., a magazine), or constructed with supplies from a hardware store (e.g., an outdoor spigot mounted on wood). Assembling these materials posed no serious or lasting obstacles.

The substantial difficulties encountered, rather, were in some cases institutional, and in all cases people-based. The purpose of the following discussion, it should be understood, is not to malign or attribute blame, but rather to allow future investigators to more thoughtfully and knowledgeably plan the nature and course of their projects.

One striking fact emerged early in the project. The freedom to attempt to implement this research project required the written approval of surprisingly many individuals: not only the principal investigator's (PI) doctoral committee and department chairman, but two human subject committees (Shands and Veterans Hospitals) with separate and different protocols, and the chiefs of each GVAH service involved (medicine, neurology, psychiatry, and nursing).

Since most of the data in this study were generated by raters, the question of who would serve as raters was one of central importance. Ideally, of course, an unlimited pool of competent and cooperative raters would have been available. Several groups were considered, including practically every care delivery discipline at every level of training to be found in a teaching hospital. The possibility of appointing our own full-time, roving rater was even

considered. In contrast to the ideal, however, it was quickly discovered that participation in such a research project was among the lowest of priorities for many of the groups approached. In certain cases, additionally, unfavorable supervisor-employee relations made it impossible for supervisors to exert influence on their subordinates to participate. Subordinates, when approached directly by the PI, occassionally tried to enmesh the PI in these conflicts by offering to participate if the PI could extract certain concessions from their supervisors. This was clearly not an appropriate role for a graduate student/researcher.

Obtaining permission to request individual medical students and residents on the involved wards to participate was particularly time-consuming. Largely at the insistence of the chairman of the GVAH human subjects committee, the PI had to secure written consent from the directors of medical education for the three GVAH services. These in turn insisted on seeing the same documentation from the dean of the University of Florida medical school, who in turn insisted on the same from the chairmen of the three departments in the medical school. All seven signatures were indeed obtained by the PI, and permission was thus granted to request the individual students' and residents' participation.

Two problems, however, remained. First, the individuals were still free to refuse participation for any reason they might have had. Naturally, this allowed any naivete, misinformation, or antipathy toward research or psychology on their part to emerge, thus presenting a formidable public relations challenge to the PI. Second, the high turnover (rotation) rate of these individuals severely limited the

usefulness of cooperative raters to the study, while also replacing the uncooperative ones.

Veterans Administration regulations prohibit staff from accepting extra remuneration for in-hospital activities. It was thus not possible to offer money to raters for participating, in either a flat-fee or piece-wise arrangement. Ultimately, the only strategy that produced real rater cooperation was the continued presence of the PI on the wards, which led to the establishment of some working personal relationships. In this context, it was possible to alternately stroke and push each rater to cooperate.

In a similar vein to the rater issues, the choice of an assistant was also a crucial one. His duties were not only to generate simultaneous ratings with the two subject-present scales, but also to take responsibility for the completion of data collection when the PI left town to do a one-year clinical internship. Time proved the choice of an individual to be a poor one, in that the assistant's reliability, commitment to the project, motivation to excel, etc., proved inadequate. His continued involvement, once his performance began to deteriorate, cost the PI considerable additional time and money, and still the assistant did not fulfill his original obligation to complete the data collection. It may be suggested that any PI who wants things done correctly should either do them himself, or establish a structure that allows him to maintain control over the reinforcers for small units of appropriate behavior by the assistant. As the present case illustrates, some formalization of the assistant's obligations is proper and desirable.

REFERENCES

Barnes, J. A. Effects of reality orientation classroom on memory loss, confusion and disorientation in geriatric patients. Gerontologist, 1974, 14 (2), 138-142.

Bender, L. A. A visual-motor gestalt test and its clinical use. New York: American Orthopsychiatric Association, 1938.

Brook, P., Degun, G., & Mather, M. Reality orientation: A therapy for psychogeriatic patients: A controlled study. British Journal of Psychiatry, 1975, 127, 42-45.

Brotman, H. B. Who are the aging? In E. W. Busse and E. Pfeiffer (Eds.), Mental illness in later life. Washington, D. C.: American Psychiatric Association, 1973.

Buros, O. K. (Ed.), The seventh mental measurements yearbook. Highland Park, N. J.: Gryphon Press, 1972.

Busse, E. W. The modern challenge of threescore and ten. Journal of the American Geriatrics Society, 1969, 17, 887-893.

Butler, R. N. Help for the elderly. American Journal of Psychiatry, 1969, 126, 725.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56 (2), 81-105.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand-McNally, 1963.

Carmichael, J., & Linn, M. W. Functioning of the elderly patient in relation to the physician's diagnosis of chronic brain syndrome. Journal of the American Geriatrics Society, 1974, 22 (5), 217-221.

Citrin, R. S., & Dixon, D. N. Reality orientation: A milieu therapy used in an institution for the aged. Gerontologist, 1977, 17 (1), 39-43.

Dastoor, D. P., Norton, S., Boillat, J., Minty, J., Papadopoulou, R., & Mueller, H. F. A psychogeriatric assessment program: I: Social functioning and ward behavior. Journal of the American Geriatrics Society, 1975, 23, 465-471.

Eisner, D. A. Can hyperbaric oxygenation improve cognitive functioning in the organically impaired elderly? A critical review. Journal of Geriatric Psychiatry, 1975, 8 (2), 173-188.

Evenson, R. Geriatric profile. St. Louis, Mo.: Missouri Institute of Psychiatry, 1971. (Unpublished)

Farina, A., Arenberg, D., & Guskin, S. A scale for measuring minimal social behavior. Journal of Consulting Psychology, 1957, 21, 265-268.

Folsom, J. C. Reality orientation for the elderly mental patient. Journal of Geriatric Psychiatry, 1968, 1 (2), 291-307.

Foster, R. W. Camelot Behavioral Checklist. Lawrence, Kansas: Camelot Behavioral Systems, 1974.

Greenwald, S. R., & Linn, M. W. The younger nursing home patient. Journal of Gerontology, 1972, 27 (3), 393-398.

Harris, C., & Ivory, P. An outcome evaluation of reality orientation therapy with geriatric patients in a state hospital. Gerontologist, 1976, 16 (6), 496-503.

Jarvik, L. F., & Milne, J. F. Gerovital-H3: A review of the literature. In S. Gershon and A. Raskin (Eds.), Aging (Vol. 2). New York: Raven Press, 1975.

Johnson, S. M., & Bolstad, O. D. Methodological issues in naturalistic observation: Some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology, concepts, and practice. Champaigne, Ill.: Research Press, 1973.

Kahn, R., Goldfarb, A., Pollack, M., & Peck, A. Brief objective measures for the determination of mental status in the aged. American Journal of Psychiatry, 1960, 117, 326-328.

Kent, R. N., O'Leary, K. D., Diament, C., & Deitz, A. Expectation biases in observational evaluation of therapeutic change. Journal of Consulting and Clinical Psychology, 1974, 42 (6), 774-780.

Kirchoff. Presentation at VA Continuing Education Course, "The aging process: Orientation vs. confusion," Birmingham, Alabama, 1977.

Kuriansky, J., & Gurland, B. The performance test of activities of daily living. International Journal of Aging and Human Development, 1976, 7 (4), 343-352.

Kuriansky, J., Gurland, B. J., Fleiss, J. L., & Cowan, D. The assessment of self-care capacity in geriatric psychiatric patients by objective and subjective methods. Journal of Clinical Psychology, 1976, 32, 95-102.

Lacks, P. B., Harrow, M., Colbert, J., & Levine, J. Further evidence concerning the diagnostic accuracy of the Halstead organic test battery. Journal of Clinical Psychology, 1970, 26, 480-481.

Lawton, M. P. The functional assessment of elderly people. Journal of the American Geriatrics Society, 1971, 19, 465-481.

Lawton, M. P. Assessing the competence of older people. In D. P. Kent, R. Kastenbaum, & S. Sherwood (Eds.), Research planning and action for the elderly. New York: Behavioral Publications, 1972.

Lawton, M. P., & Brody, E. Assessment of older people: Self-maintaining instrumental activities of daily living. Gerontologist, 1969, 9, 179-186.

Letcher, P. B., Peterson, L. P., & Scarbrough, D. Reality orientation: A historical study of patient progress. Hospital and Community Psychiatry, 1974, 25 (12), 801-803.

Lindsley, O. R. Geriatric behavioral prosthetics. In R. Kastenbaum (Ed.), New thoughts on old age. New York: Springer, 1964.

Linn, M. W. A rapid disability rating scale. Journal of the American Geriatrics Society, 1967, 15, 211-214.

Linn, M. W., & Gurel, L. Initial reactions to nursing home placement. Journal of the American Geriatrics Society, 1969, 17 (2), 219-223.

Linn, M. W., Gurel, L., & Linn, B. S. Patient outcome as a measure of quality of nursing home care. American Journal of Public Health, 1977, 67 (4), 337-342.

List of research and evaluation tools. Tuscaloosa, Alabama: VA Hospital, 1976.

Lowenthal, M. F. Lives in distress. New York: Basic Books, 1964.

Mangun, C. W, & Webb, W. W. Promazine hydrochoride in the treatment of chronic catatonic schizophrenics. Journal of Nervous and Mental Diseases, 1956, 123, 553-556.

Markson, E. W., & Levitz, G. A. A Guttman scale to assess memory loss among the elderly. Gerontologist, 1973, 13, 337-340.

Meehl, P. E., & Rosen, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. Psychological Bulletin, 1955, 52, 194-216.

Meer, B., & Baker, J. A. The Stockton Geriatric Rating Scale. Journal of Gerontology, 1966, 21, 392-403.

Miller, E. R., & Parachek, J. F. Validation and standardization of a goal-oriented, quick screening geriatric scale. Journal of the American Geriatrics Society, 1974, 22, 278-283.

Neidtich, J. A., & White, L. Prediction of short-term outcome in newly admitted psychogeriatric patients. Journal of the American Geriatrics Society, 1976, 24 (2), 72-78.

Ogren, E. H., & Linn, M. W. Male nursing home patients: Relocation and mortality. Journal of the American Geriatrics Society, 1971, 19 (3), 229-239.

O'Leary, K. D., Kent, R. N., Kanowitz, J. Shaping data collection with experimental hypotheses. Journal of Applied Behavior Analysis, 1975, 8 (1), 43-51.

Plutchik, R., & Conte, H. Change in social and physical functioning of geriatric patients over a one-year period. Gerontologist, 1972, 12, 181-184.

Plutchik, R., Conte, H., Lieberman, M., Bakur, M., Grossman, J., & Lehrman, N. Reliability and validity of a scale for assessing the functioning of geriatric patients. Journal of the American Geriatrics Society, 1970, 18 (6), 491-500.

Rebok, G. W., & Hoyer, W. J. The functional context of elderly behavior. Gerontologist, 1977, 17, 27-34.

Redick, R. W., Kramer, M., & Taube, C. A. Epidemiology of mental illness and utilization of psychiatric facilities among older persons. In E. W. Busse & E. Pfeiffer (Eds.), Mental illness in later life. Washington, D. C.: American Psychiatric Association, 1973.

Repp, A. C., Deitz, D. E., Boles, S. M., Deitz, S. M., & Repp, C. R. Differences among common methods for calculating interobserver agreement. Journal of Applied Behavior Analysis, 1976, 9, 109-113.

Romanczyk, R. G., Kent, R. N., Diament, C., & O'Leary, K. D. Measuring the reliability of observational data: A reactive process. Journal of Applied Behavior Analysis, 1973, 6 (1), 175-184.

Salzman, C., Kochansky, G. E., & Shader, R. I. Rating scales for geriatric psychopharmacology: A review. Psychopharmacology Bulletin, 1972, 8 (3), 3-50.

Schaie, K. W. Quasi-experimental research designs in the psychology of aging. In J. E. Birren & K. W. Schale (Eds.), Handbook of the Psychology of Aging. New York: Van Nostrand Reinhold, 1977.

Schuller, D. Y., & McNamara, J. R. Expectancy factors in behavioral observation. Behavior Therapy, 1976, 7, 519-527.

Smith, J. M., Bright, B., & McCloskey, J. Factor analytic composition of the GRS. Journal of Gerontology, 1977, 32, 58-62.

Standards for educational and psychological tests. Washington, D. C.: American Psychological Association, 1974.

Stephens, L. P. (Ed.), Reality orientation. Washington, D. C.,: American Psychiatric Association, 1970.

Thompson, L. W. Effects of hyperbaric oxygenation on behavioral functioning in elderly persons with intellectual impairment. In S. Gershon & A. Raskin (Eds.), Aging (Vol. 2). New York: Raven Press, 1975.

Walls, R. T., Werner, T. J., Bacon, A., & Zane, T. Behavior checklist. In J. Cone & R. Hawkins (Eds.), Behavior assessment: New directions in clinical psychology. New York: Brunner-Mazel, 1977.

Whanger, A. D. Geriatric mental health in North Carolina. North Carolina Journal of Mental Health, 1971, 5, 43-49.

BIOGRAPHICAL SKETCH

Paul Kenneth Chafetz was born April 22, 1953, in Memphis, Tennessee, where he graduated from Christian Brothers High School in 1971. His undergraduate career included a year of study at the University of Freiberg, Freiberg, West Germany, and he received his Bachelor of Arts from Brown University in 1975. He entered the University of Florida as a graduate student in clinical psychology later that year, and was admitted to doctoral candidacy in November, 1978. Mr. Chafetz completed a one year clinical psychology internship at Duke University Medical Center, Durham, North Carolina, in August, 1980. He currently resides in Houston, Texas, where he is a fellow in the clinical geropsychology program at the Texas Research Institute of Mental Sciences.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Walter Cunningham, Chairman
Associate Professor of Clinical Psychology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Warren Rice
Associate Professor of Clinical Psychology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Suzanne Johnson
Assistant Professor of Clinical Psychology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Cynthia Belar
Associate Professor of Clinical Psychology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Gordon Streib
Graduate Research Professor of Sociology

This dissertation was submitted to the Graduate Faculty of the College of Health Related Professions and to the Graduate Council, and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

March, 1981

Dean, College of Health Related Professions

Dean, Graduate School